DOCUMENT RESUME

ED 288 887	TM 870 616

AUTHOR	Beaton, Albert E.
TITLE	Implementing the New Design: The NAEP 1983-84 Technical Report.
INSTITUTION	National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	Center for Statistics (OERI/ED), Washington, DC.
REPORT NO	ISBN-0-88685-062-2; NAEP-15-TR-20
PUB DATE	Mar 87
GRANT	NIE-G-83-0011
NOTE	813p.; For the "Users' Guide" and "Codebooks and Layouts" pertaining to the NAEP Public-Use data tapes for 1983-84, see TM 870 621-624.
AVAILABLE FROM	National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, CN 6710, Princeton, NJ 08541-6710 ($25.00, plus $3.00 postage).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF05/PC33 Plus Postage.
DESCRIPTORS	Academic Achievement; Data Analysis; Data Collection; Data Interpretation; *Educational Assessment; Elementary Secondary Education; *Item Sampling; Latent Trait Theory; *National Competency Tests; National Surveys; Program Implementation; Reading Tests; *Research Design; Sampling; Scaling; Statistical Analysis; *Testing Programs; Writing Evaluation
IDENTIFIERS	*Balanced Incomplete Block Spiralling; *National Assessment of Educational Progress

ABSTRACT

In 1982, the Educational Testing Service (ETS) proposed to implement a new, complex design for the National Assessment of Educational Progress (NAEP). The major features of this design are described in "A New Design for a New Era" (Messick, Beaton, and Lord, 1983). The purpose of this document is to describe the actual implementation of the design in the 1983-84 National Assessment of Reading and Writing (NAEP's fifteenth year); it is intended as a supplement to the reports of that assessment (see ED 264 550, ED 273 680, ED 273 994) and supports these reports by providing detailed technical information so that the accuracy of the substantive results can be judged. Some major features of the new design were: to sample grades 4, 8, and 11 as well as students' ages 9, 13, and 17 (in school); to introduce Balanced Incomplete Block (BIB) spiralling as a method of estimating inter-relationships among variables; to collect extensive information about teachers, principals, and schools; and to scale the reading data, if possible. These innovations were added to the previously used procedures, which were kept to ensure maintenance of NAEP trends. This report describes: (1) the data collection processes, including the assessment instruments for reading and writing; (2) the data analysis process for both reading and writing, including "plausible values" of reading proficiency and the NAEP reading and writing; and (3) some estimates of the reading and writing proficiencies of selected subpopulations of the sampled students. Two supplementary studies on the validity of NAEP's reading and writing assessment instruments and the design effects in the 1983-84 sample are also presented. A glossary of terms and a 124-item reference list complete the document. (JGL)

ED288887

IMPLEMENTING THE NEW DESIGN:

THE NAEP 1983-84 TECHNICAL REPORT

Albert E. Beaton

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E. Driscoll

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

REPORT NO: 15-TR-20

THE NATION'S REPORT CARD
naep

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Assessment of Educational Progress

ETS®

BEST COPY AVAILABLE

IMPLEMENTING THE NEW DESIGN:

THE NAEP 1983-84 TECHNICAL REPORT

Albert E. Beaton

in collaboration with

John L. Barone, Anne Campbell, John J. Ferris,
David S. Freund, Eugene G. Johnson, Janet R. Johnson,
Bruce A. Kaplan, Debra L. Kline, Robert J. Mislevy,
Ina V. S. Mullis, Norma A. Norris, Alfred M. Rogers,
Kathleen M. Sheehan, Marilyn Wingersky, Rebecca Zwick

Educational Testing Service • Princeton, NJ

and

John Burke, Nancy Caldwell, Morris H. Hansen,
Josefina A. Lago, Renee Slobasky, Benjamin J. Tepping

Westat, Inc. • Washington, DC

March 1987

THE NATION'S REPORT CARD
naep

National Assessment of Educational Progress

ETS

The National Assessment of Educational Progress is funded by the U.S. Department of Education under a grant to Educational Testing Service. National Assessment is an education research project mandated by Congress to collect data over time on the performance of young Americans in various learning areas. It makes available information on assessment procedures to state and local education agencies.

This report, No. 15-TR-20, can be ordered from the National Assessment of Educational Progress at Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541.

Library of Congress Catalog Card Number 87-60432

ISBN 0-88685-062-2

The work upon which this publication is based was performed pursuant to Grant No. NIE-G-83-0011 of the Office for Educational Research and Improvement, Center for Statistics. It does not, however, necessarily reflect the views of that agency.

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, ETS, and ETS are registered trademarks of Educational Testing Service.

IMPLEMENTING THE NEW DESIGN:
THE NAEP 1983-84 TECHNICAL REPORT

CONTENTS

PAGE

Index of Tables and Figures. vii
Executive Summary. xv
Acknowledgments. xvii

PART I

Chapter 1 Introduction to the Technical Report 3
Albert E. Beaton

Chapter 2 Overview of Part I: The Design and Implementation of the Year 15 NAEP (and tabular summary of NAEP data). . 13
Albert E. Beaton

Chapter 3 Development of the Year 15 NAEP Reading and Writing Assessments. 47
Ina V. S. Mullis

Chapter 4 Sample Selection and Instrument Collection 79
Morris H. Hansen, Benjamin J. Tepping, Josefina A. Lago, John Burke

Chapter 5 The Assignment of Exercises to Students. 97
Albert E. Beaton, Eugene G. Johnson, John J. Ferris

Chapter 6 Instrument and Item Information. 119
Janet R. Johnson

Chapter 7 Field Administration 135
Renee Slobasky, Nancy Caldwell

Chapter 8 Materials Processing and Database Creation 161
John L. Barone

Chapter 8.1 Processing Assessment Materials. 165
Alfred M. Rogers, Norma A. Norris

Chapter 8.2 Professional Scoring 175
Anne Campbell

PAGE

Chapter 8.3 Data Entry System. 185
Alfred M. Rogers

Chapter 8.4 Editing Data 201
Alfred M. Rogers

Chapter 8.5 Quality Control. 205
John J. Ferris

Chapter 8.6 Database Creation. 211
Alfred M. Rogers

Chapter 8.7 Public-Use Data Tape Construction. 215
Alfred M. Rogers

PART II

Chapter 9 Overview of Part II: Analysis of the Year 15 NAEP Data. 225
Albert E. Beaton

Chapter 10 The Reading Data Analysis: Introduction 239
Robert J. Mislevy

Chapter 10.1 Assessment of the Dimensionality of Year 15 Reading Data 245
Rebecca Zwick

Chapter 10.2 Joint Estimation Procedures. 285
Marilyn Wingersky, Bruce A. Kaplan,
Albert E. Beaton

Chapter 10.3 Marginal Estimation Procedures 293
Robert J. Mislevy, Kathleen M. Sheehan

Chapter 10.4 Trend Analysis 361
Robert J. Mislevy, Kathleen M. Sheehan

Chapter 10.5 The NAEP Reading Scale 381
Albert E. Beaton

Chapter 11 The Writing Data Analysis: Introduction 391
Albert E. Beaton

Chapter 11.1 The Writing Exercise Data. 397
Albert E. Beaton

Chapter 11.2 The Effect of Mode of Item Administration (BIB Spiral or Paced Tape) on Estimates of Writing Performance . . 405
Eugene G. Johnson

PAGE

Chapter 11.3 Estimation of Trends in Writing Achievement. 431
Eugene G. Johnson

Chapter 11.4 The Average Response Method (ARM) of Scaling 435
Albert E. Beaton, Eugene G. Johnson

Chapter 12 Background and Attitude Data Analysis. 481
Albert E. Beaton, Norma A. Norris,
Janet R. Johnson

Chapter 13 Parameter Estimation 491
Albert E. Beaton

Chapter 13.1 Weighting Procedures 493
Eugene G. Johnson, Morris H. Hansen,
Benjamin J. Tepping, Josefina A. Lago, John Burke

Chapter 13.2 Estimation of Uncertainty Due to Sampling Variability 505
Eugene G. Johnson

Chapter 13.3 Estimation of Variability Due to Imputation. 513
Robert J. Mislevy

Chapter 13.4 Use of the NAEP Almanacs 517
Rebecca Zwick

Chapter 14 Supplementary Studies. 523
Albert E. Beaton

Chapter 14.1 Validity Issues in NAEP: Year 15 Reading and Writing Assessments. 525
Rebecca Zwick

Chapter 14.2 Design Effects 545
Eugene G. Johnson

PART III

Chapter 15 Estimates of the Reading and Writing Proficiency of American Students 565
Albert E. Beaton, David S. Freund,
Bruce A. Kaplan

Appendix A: Assessment Items 645
Appendix B: Reading Trend Analysis Items 675
Glossary 735
List of References 751
Subject Index. 763

INDEX OF TABLES AND FIGURES

NUMBER	TITLE	PAGE
2(1)	NAEP learning areas, grades and ages assessed.	16
2(2)	Measurement instruments developed by ETS	29
2(3)	Number of items administered	30
2(4)	Number of reading and writing exercises by type of administration	31
2(5)	Allocation of PSUs to regions and community types. . . .	32
2(6)	Characteristics of schools	33
2(7)	Number of responses to teacher questionnaire	35
2(8)	Number of assessment sessions by administration type . .	36
2(9)	Number of students by administration type.	37
2(10)	Spiral sample by demographic characteristics, Grade 4/Age 9.	38
2(11)	Spiral sample by demographic characteristics, Grade 8/Age 13	39
2(12)	Spiral sample by demographic characteristics, Grade 11/Age 17.	40
2(13)	Excluded student sample by demographic characteristics, Grade 4/Age 9	41
2(14)	Excluded student sample by demographic characteristics, Grade 8/Age 13.	42
2(15)	Excluded student sample by demographic characteristics, Grade 11/Age 17	43
2(16)	Tape sample by demographic characteristics, Grade 4/Age 9.	44
2(17)	Tape sample by demographic characteristics, Grade 8/Age 13	45
2(18)	Tape sample by demographic characteristics, Grade 11/Age 17.	46
4(1)	Summary of school participation experience	87
4(2)	Weighted and unweighted distribution of excluded students, by reason for exclusion and grade/age. . . .	91
4(3)	Comparisons of Year 15 target assessments to actual assessments, by grade/age	92
4(4)	Comparison of Year 15 and Year 13 proportion of excluded students, by grade/age.	93
4(5)	Comparison of Years 15 and Year 13 student participation rates, by type of PSU and grade/age. . .	94
4(6)	Distribution of teachers by age class and participation status	96

NUMBER	TITLE	PAGE
5(1)	Sample design summary	101
5(2)	Booklet design, BIB spiral sample	102
5(3)	Spiral sample, block-to-booklet correspondence	105
5(4)	Booklet design, UBIB spiral sample	106
5(5)	Number of pairings of item blocks in spiral design	107
5(6)	Number of booklets administered	111
5-1	BIB spiral sample, number of students per block	113
5(7)	Number of blocks administered, spiral and tape samples	114
5-2	BIB spiral, UBIB spiral and tape samples: Number of students per booklet	115
6(1)	Booklet contents by block, Grade 4/Age 9	120
6(2)	Booklet contents by block, Grade 8/Age 13 and Grade 11/Age 17	121
6(3)	Assessment items for Grade 4/Age 9	123
6(4)	Assessment items for Grade 8/Age 13	124
6(5)	Assessment items for Grade 11/Age 17	125
6(6)	Items by block on tapes, Age 9	128
6(7)	Items by block on tapes, Age 13	129
6(8)	Items by block on tapes, Age 17	130
6(9)	Year 15 writing items	131
7(1)	Criteria met by NAEP supervisors, by region	138
7(2)	Frequency of makeup sessions	147
7(3)	Regular and makeup sessions conducted	147
7(4)	Change in attendance rates with makeup sessions	147
7(5)	Number and distribution of quality control visits	153
8-1	NAEP data flow overview	162
8.2(1)	Distribution of reading and writing exercises	178
8.3-1	Student data entry processing	186
10.1(1)	Number of items and students available for dimensionality analyses	252
10.1(2)	Eigenvalues and descriptive statistics for phi matrices	254
10.1(3)	Eigenvalues and descriptive statistics for tetrachoric matrices	255
10.1(4)	Eigenvalues of the image correlation matrix	260
10.1(5)	First five eigenvalues of correlation and image correlation matrices for simulation data	261
10.1(6)	Results of Rosenbaum analyses	267
10.1(7)	Subjects available to estimate within- and across-block correlations for 30-item BIB simulation	270
10.1(8)	Distribution of residual correlations for 30-item BIB simulation	271

NUMBER	TITLE	PAGE
10.1(9)	Partial comparison of eigenstructure of BIB and complete data correlation matrices for 30-item simulation . . .	272
10.2(1)	Values assigned to examinees whose maximum likelihood estimates could not be computed	290
10.2(2)	Minimum and maximum scores and number of values changed by grade/age	290
10.2-1	Distributions of adjusted proficiency scale scores . . .	292
10.3(1)	Blocks selected for scaling the Year 15 reading data . .	301
10.3-1	Diagnostic fit plot for item 9	303
10.3-2	Bias plot for item 10.	303
10.3(2)	Coding of background variables, BIB data	305
10.3(3)	Estimated conditional effects, BIB data.	307
10.3(4)	Sampling procedure used to generate plausible values . .	309
10.3(5)	Reliability coefficients by booklet, Grade 4/Age 9 . . .	317
10.3(6)	Reliability coefficients by booklet, Grade 8/Age 13. . .	319
10.3(7)	Reliability coefficients by booklet, Grade 11/Age 17 . .	321
10.3(8)	Approximate shrinkage of regression coefficients of nonconditioned background variables, Grade 4/Age 9 . .	324
10.3(9)	Approximate shrinkage of regression coefficients of nonconditioned background variables, Grade 8/Age 13. .	325
10.3(10)	Approximate shrinkage of regression coefficients of nonconditioned background variables, Grade 11/Age 17 .	326
10.3-3	BIB-Pace percent correct, Age 9 total, IRT items	328
10.3-4	BIB-Pace percent correct, Age 9 male, IRT items.	329
10.3-5	BIB-Pace percent correct, Age 9 female, IRT items. . . .	330
10.3-6	BIB-Pace percent correct, Age 9 white, IRT items	331
10.3-7	BIB-Pace percent correct, Age 9 black, IRT items	332
10.3-8	BIB-Pace percent correct, Age 9 Hispanic, IRT items. . .	333
10.3-9	BIB-Pace percent correct, Age 13 total, IRT items. . . .	334
10.3-10	BIB-Pace percent correct, Age 13 male, IRT items	335
10.3-11	BIB-Pace percent correct, Age 13 female, IRT items . . .	336
10.3-12	BIB-Pace percent correct, Age 13 white, IRT items. . . .	337
10.3-13	BIB-Pace percent correct, Age 13 black, IRT items. . . .	338
10.3-14	BIB-Pace percent correct, Age 13 Hispanic, IRT items . .	339
10.3-15	BIB-Pace percent correct, Age 17 total, IRT items. . . .	340
10.3-16	BIB-Pace percent correct, Age 17 male, IRT items	341
10.3-17	BIB-Pace percent correct, Age 17 female, IRT items . . .	342
10.3-18	BIB-Pace percent correct, Age 17 white, IRT items. . . .	343
10.3-19	BIB-Pace percent correct, Age 17 black, IRT items. . . .	344
10.3-20	BIB-Pace percent correct, Age 17 Hispanic, IRT items . .	345
10.3(11)	Correlations and regression coefficients for spiral vs. paced percents correct of IRT reading items. . . .	346
10.3-21	BIB/Pace population equating, Age 9.	351
10.3-22	BIB/Pace population equating, Age 13	352
10.3-23	BIB/Pace population equating, Age 17	353
10.3(12)	BIB/Pace subgroup means, Age 9	354

NUMBER	TITLE	PAGE
10.3(13)	BIB/Pace subgroup means, Age 13.	355
10.3(14)	BIB/Pace subgroup means, Age 17.	357
10.3-24	Comparison of estimated b values, Age 9 vs. 13	359
10.3-25	Comparison of estimated b values, Age 17 vs. 13.	360
10.4(1)	Booklets selected for calibrating trend items.	364
10.4(2)	"Changed" reading items.	365
10.4(3)	Proportion correct for questionable items.	366
10.4-1	Diagnostic plot for item 87.	368
10.4-2	Plots of items excluded during preliminary calibrations.	369
10.4(4)	Items excluded during preliminary calibration runs . . .	372
10.4(5)	Item calibration summary	372
10.4(6)	Additional booklets used for estimating conditional distributions.	375
10.4(7)	Estimated conditional effects, Year 2 Pace data.	376
10.4(8)	Estimated conditional effects, Year 6 Pace data.	377
10.4(9)	Estimated conditional effects, Year 11 Pace data	378
10.4(10)	Estimated conditional effects, Year 15 Pace data	379
10.5-1	Levels of proficiency.	389
11(1)	Year 15 writing items.	394
11.1(1)	Percentages of exact score point agreement and intra-class correlation coefficients for primary trait scoring, Year 15	398
11.1(2)	Reliability statistics for primary trait ratings	400
11.1(3)	Percentages of exact score point agreement and intra-class correlation coefficients for primary trait scoring conducted in 1983-84	401
11.1(4)	Number of students responding to each writing exercise .	403
11.2(1)	Writing exercises selected for the BIB/Pace comparison .	409
11.2(2)	Effect of mode of administration on writing performance, Age 9, "Aunt May"	410
11.2(3)	Effect of mode of administration on writing performance, Age 9, "Dali"	412
11.2(4)	Effect of mode of administration on writing performance, Age 9, "Hole in the Box".	414
11.2(5)	Effect of mode of administration on writing performance, Age 13, "Split Session"	416
11.2(6)	Effect of mode of administration on writing performance, Age 13, "Dali".	418
11.2(7)	Effect of mode of administration on writing performance, Age 13, "Hole in the Box"	420
11.2(8)	Effect of mode of administration on writing performance, Age 17, "Split Session"	422
11.2(9)	Effect of mode of administration on writing performance, Age 17, "Dali".	424

NUMBER	TITLE	PAGE
11.2(10)	Effect of mode of administration on writing performance, Age 17, "Hole in the Box"	426
11.2-1	Difference between BIB and Pace percentages, Age 9 . . .	428
11.2-2	Difference between BIB and Pace percentages, Age 13. . .	429
11.2-3	Difference between BIB and Pace percentages, Age 17. . .	430
11.3(1)	Exercises used to estimate trends in writing performance	433
11.4(1)	Writing items for the ARM writing scale.	450
11.4(2)	Distribution by grade of the number of writing scale items taken by a student	450
11.4-1a	Grade 4 conditioned variables, group effects	465
11.4-1b	Grade 8 conditioned variables, group effects	466
11.4-1c	Grade 11 conditioned variables, group effects.	467
11.4-2a	Grade 4 unconditioned variables, group effects	468
11.4-2b	Grade 8 unconditioned variables, group effects	469
11.4-2c	Grade 11 unconditioned variables, group effects.	470
11.4-3a	Grade 4 F-values converted to N(0,1)	475
11.4-3b	Grade 8 F-values converted to N(0,1)	476
11.4-3c	Grade 11 F-values converted to N(0,1).	477
11.4-4a	Grade 4 SE (ARM) vs. SE (Meanparts).	478
11.4-4b	Grade 8 SE (ARM) vs. SE (Meanparts).	479
11.4-4c	Grade 11 SE (ARM) vs. SE (Meanparts)	480
12(1)	Reporting subgroup variables	482
12(2)	Race/ethnicity classifications	485
12(3)	Determining race/ethnicity	486
12(4)	Geographic regions	488
13.1(1)	Major subgroups for post-stratification.	503
13.4(1)	Year 15 almanacs, dates of issue and comments.	518
14.1(1)	Reading and writing (ARM) proficiency means for selected groups.	530
14.1(2)	Correlations of reading, writing, and selected background variables	535
14.1(3)	Correlations of reading, writing, PSAT scores, and selected background variables.	536
14.1(4)	Definition of variables for analyses	537
14.1(5)	Correlations of reading and writing with frequency of reading and writing activities.	538
14.1(6)	Item text and response codes for reading and writing . .	539
14.2(1)	Distributions of design effects for demographic subgroups, Grade 4	551
14.2-1a	Grade 4, log base 10 of design effects	552
14.2-1b	Grade 4, log base 10 of design effects	553

NUMBER	TITLE	PAGE
14.2-1c	Grade 4, log base 10 design effects.	554
14.2(2)	Distributions of design effects for demographic subgroups, Grade 8	555
14.2-2a	Grade 8, log base 10 of design effects	556
14.2-2b	Grade 8, log base 10 of design effects	557
14.2-2c	Grade 8, log base 10 of design effects	558
14.2(3)	Distributions of design effects for demographic subgroups, Grade 11.	559
14.2-3a	Grade 11, log base 10 of design effects.	560
14.2-3b	Grade 11, log base 10 of design effects.	561
14.2-3c	Grade 11, log base 10 of design effects.	562
15(1)	Number of students by grade/age combination and by type of assessment.	571
15(2)	Number of spiral students by grade/age	572
15(3)	Estimated total number of students in the population eligible for spiral assessment, Grade 4/Age 9.	575
15(4)	Estimated total number of students in the population eligible for spiral assessment, Grade 8/Age 13	576
15(5)	Estimated total number of students in the population eligible for spiral assessment, Grade 11/Age 17. . . .	577
15(6)	Estimated total number of students in the population eligible for spiral assessment who would be deemed unassessable by their schools, Grade 4/Age 9	578
15(7)	Estimated total number of students in the population eligible for spiral assessment who would be deemed unassessable by their schools, Grade 8/Age 13.	579
15(8)	Estimated total number of students in the population eligible for spiral assessment who would be deemed unassessable by their schools, Grade 11/Age 17	580
15(9)	Estimated total number of students eligible for assessment by tape sample, Grade 4/Age 9	581
15(10)	Estimated total number of students eligible for assessment by tape sample, Grade 8/Age 13.	582
15(11)	Estimated total number of students eligible for assessment by tape sample, Grade 11/Age 17	583
15(12)a	Number of students receiving reading and writing items and plausible values, Grade 4/Age 9.	584
15(12)b	Weighted counts of students receiving reading and writing items and plausible values, Grade 4/Age 9. . .	585
15(13)a	Number of students receiving reading and writing items and plausible values, Grade 8/Age 13	586
15(13)b	Weighted counts of students receiving reading and writing items and plausible values, Grade 8/Age 13 . .	587
15(14)a	Number of students receiving reading and writing items and plausible values, Grade 11/Age 17.	588

NUMBER	TITLE	PAGE
15(14)b	Weighted counts of students receiving reading and writing items and plausible values, Grade 11/Age 17. .	589
15(15)	General reading proficiency estimates, Grade 4	590
15(16)	General reading proficiency estimates, Grade 8	591
15(17)	General reading proficiency estimates, Grade 11.	592
15(18)	ARM writing proficiency estimates, Grade 4	593
15(19)	ARM writing proficiency estimates, Grade 8	594
15(20)	ARM writing proficiency estimates, Grade 11.	595
15(21)	Reading proficiency, imputed student grade, Grade 4. . .	596
15(22)	Reading proficiency, student sex, Grade 4.	597
15(23)	Reading proficiency, ethnicity/race, Grade 4	598
15(24)	Reading proficiency, region, Grade 4	599
15(25)	Reading proficiency, imputed student age, Grade 4. . . .	600
15(26)	Reading proficiency, size/type of community, Grade 4 . .	601
15(27)	Reading proficiency, parental education, Grade 4	602
15(28)	Reading proficiency, % at or above anchor, Grade 4 . . .	603
15(29)	Reading proficiency, imputed student grade, Grade 8. . .	604
15(30)	Reading proficiency, student sex, Grade 8.	605
15(31)	Reading proficiency, ethnicity/race, Grade 8	606
15(32)	Reading proficiency, region, Grade 8	607
15(33)	Reading proficiency, imputed student age, Grade 8. . . .	608
15(34)	Reading proficiency, size/type of community, Grade 8 . .	609
15(35)	Reading proficiency, parental education, Grade 8	610
15(36)	Reading proficiency, % at or above anchor, Grade 8 . . .	611
15(37)	Reading proficiency, imputed student grade, Grade 11 . .	612
15(38)	Reading proficiency, student sex, Grade 11	613
15(39)	Reading proficiency, ethnicity/race, Grade 11.	614
15(40)	Reading proficiency, region, Grade 11.	615
15(41)	Reading proficiency, imputed student age, Grade 11 . . .	616
15(42)	Reading proficiency, size/type of community, Grade 11. .	617
15(43)	Reading proficiency, parental education, Grade 11. . . .	618
15(44)	Reading proficiency, % at or above anchor, Grade 11. . .	619
15(45)	Writing proficiency, imputed student grade, Grade 4. . .	620
15(46)	Writing proficiency, student sex, Grade 4.	621
15(47)	Writing proficiency, ethnicity/race, Grade 4	622
15(48)	Writing proficiency, region, Grade 4	623
15(49)	Writing proficiency, imputed student age, Grade 4. . . .	624
15(50)	Writing proficiency, size/type of community, Grade 4 . .	625
15(51)	Writing proficiency, parental education, Grade 4	626
15(52)	Writing proficiency, % at or above anchor, Grade 4 . . .	627
15(53)	Writing proficiency, imputed student grade, Grade 8. . .	628
15(54)	Writing proficiency, student sex, Grade 8.	629

NUMBER	TITLE	PAGE
15(55)	Writing proficiency, ethnicity/race, Grade 8	630
15(56)	Writing proficiency, region, Grade 8	631
15(57)	Writing proficiency, imputed student age, Grade 8	632
15(58)	Writing proficiency, size/type of community, Grade 8	633
15(59)	Writing proficiency, parental education, Grade 8	634
15(60)	Writing proficiency, % at or above anchor, Grade 8	635
15(61)	Writing proficiency, imputed student grade, Grade 11	636
15(62)	Writing proficiency, student sex, Grade 11	637
15(63)	Writing proficiency, ethnicity/race, Grade 11	638
15(64)	Writing proficiency, region, Grade 11	639
15(65)	Writing proficiency, imputed student age, Grade 11	640
15(66)	Writing proficiency, size/type of community, Grade 11	641
15(67)	Writing proficiency, parental education, Grade 11	642
15(68)	Writing proficiency, % at or above anchor, Grade 11	643

IMPLEMENTING THE NEW DESIGN:
THE NAEP 1983-84 TECHNICAL REPORT

EXECUTIVE SUMMARY

In 1982, Educational Testing Service (ETS) proposed to implement a new, complex design for the National Assessment of Educational Progress (NAEP). The major features of the design are described in A New Design for A New Era (Messick, Beaton, & Lord, 1983). ETS received the NAEP grant on July 1, 1983. The purpose of this technical report is to document the implementation of the design in the 1983-84 national assessment of reading and writing.

This is a technical report describing assessment processes; it is *not* intended to report and interpret the performance of students at various grade and age levels. Such results are presented in the NAEP reports *The Reading Report Card: Progress Toward Excellence in Our Schools* (1985), *Writing: Trends Across the Decade, 1974-84* (Applebee, Langer, & Mullis, 1986a), and *The Writing Report Card: Writing Achievement in American Schools, 1984* (Applebee, Langer, & Mullis, 1986b). This report supports those reports by providing detailed technical information so that the adequacy of the substantive results can be judged.

The introduction to this report presents the major features of the ETS design and shows how ETS met or exceeded its commitments. Some major features were: to sample grades 4, 8, and 11 as well as 9-, 13-, and in-school 17-year-olds; to introduce BIB spiralling as a method of estimating inter-relationships among variables; to collect extensive information about teachers, principals, and schools; and to scale the reading data, if possible. These innovations were to be introduced in addition to the previously used procedures which were kept to ensure the maintenance of NAEP trends.

The rest of the technical report is divided into three parts. Part I describes the processes involved in collecting the NAEP data. The story begins with the development of the NAEP data assessment instruments for reading and writing. A representative sample of American students in public and private schools was selected by Westat, Inc., the ETS subcontractor for sampling and field administration. BIB spiralling was used to assign assessment exercises to students to allow the study of inter-relationships of assessment exercises within a subject area, such as reading or writing, and between the subject areas and the background and attitude questions. Westat's field administration, from contacting the schools through checking for completeness of data, is detailed. The processes involved in converting the responses of students from their assessment booklets to a carefully edited database is discussed. Finally, the quality control checks are described.

Part II describes the data analyses. The analysis of the reading data is reported first. The dimensionality of the reading data was explored, and no reason was found to reject the assumption of unidimensionality for the exercises used in the reading scale. Using item response theory, the item parameters of the reading exercises were estimated, and individual student distributions of plausible values for reading proficiency were created. Plausible values are a device for encoding both what we know and what we do not know about an individual's proficiency. A single scale, linking the three ages and grades, was developed; this scale was linked to data collected in past assessments, back to the 1970-71 school year. The NAEP reading scale was behaviorally anchored to enable succinct reporting of what the population of students can and cannot do.

A NAEP writing scale was also produced by the development of a scaling procedure called the Average Response Method. This method estimates how students would have performed if they had been administered a particular set of ten writing exercises. The NAEP writing scale was applied to the students in the fourth, eighth, and eleventh grades only. A study of the effect of changing the way in which NAEP was administered showed substantial differences in response patterns and so the data from past assessments were not merged with the BIB spiralled data. For the writing trend report, the trends were maintained using the "bridge" data, which were collected using a tape recorder and reported using average percentages correct, as in past writing reports.

A method for scaling background and attitude questions was also developed, but has not yet been used extensively.

The processes involved in estimating the performances of populations of students is described next in the report. These processes include the computation of sampling weights, the estimation of sampling error using the jackknife method, and the estimation of variability due to imputation. Finally, the use of the NAEP's many tables of parameter estimates is described.

Two supplementary studies are also presented: the first is a study of the validity of the NAEP reading and writing assessment instruments; the second is a study of the design effects in the 1983-84 sample.

Part III presents some estimates of the reading and writing proficiencies of the sampled populations of students. Many thousands of pages of such tables have been developed, far too many to include in this report. This part of the report presents a few selected tables which estimate the proficiencies of important subpopulations, such as the different genders, racial/ethnic groupings, and regions of the country.

The NAEP 1983-84 data, as well as data from all past assessments, are available on public-use data tapes. All student, teacher, principal, and school data are available, except for a few items which might compromise the confidentiality of the respondents. The plausible values for reading and writing are also available on the tape.

ACKNOWLEDGMENTS

The production of this report, as well as the data and data analyses that it documents, is the result of the contributions of many persons, at ETS and elsewhere. It is fitting that these individuals be acknowledged and receive my thanks for their contributions to NAEP.

I wish to thank the ETS management for its support, especially Greg Anrig, Bob Solomon, and Henry Braun, and also the NAEP project management, Archie Lapointe, Ina Mullis, Doug Rhodes, and Paul Barton.

I also wish to thank the Assessment Policy Committee, with Wilmer Cody as Chair, which has contributed its wisdom and support.

I am also grateful to Sam Messick, who was primarily responsible for the creation of the new NAEP design. Many other persons on the ETS and Westat staffs have made substantial contributions. Fred Lord must be singled out for his major role.

Ina Mullis has had a major influence upon our evolving conception of NAEP. Her primary responsibility has been for the development of the measurement instruments and the interpretation of results.

Jules Goodison was primarily responsible for the operational aspects of the 1983-84 NAEP and has been a major contributor to the implementation of the design.

Kent Ashworth has been particularly helpful by encouraging the publication of this and other reports.

The Technical Advisory Committee, which has maintained oversight of the NAEP design and analysis, has been outstanding. The members are Robert Linn, University of Illinois (chair); Sylvia Johnson, Howard University (vice-chair); Robert Glaser, University of Pittsburgh; Bert Green, Johns Hopkins University; Melvin Novick, University of Iowa; Richard Snow, Stanford University; and Ingram Olkin, Stanford University, who replaced Melvin Novick in 1986. Barbara Shapiro served as liaison between the Assessment Policy Committee and the Technical Advisory Committee.

* * *

The implementation of the data analysis was the function of the Psychometric Research Department, of which I am director. The staff members made many creative contributions to the statistical and psychometric

aspects of NAEP, as well as managing the day-to-day operation of the analyses. I am particularly grateful for the contributions of Janet Johnson, Gene Johnson, Fred Lord, Bob Mislevy, Kathy Sheehan, Marilyn Wingersky, and Rebecca Zwick.

I am also very grateful to the Data Analysis Department, under the consummate direction of John Barone, for developing the operating systems and producing the data analyses. Developing and operating the NAEP data management systems under severe time constraints were the province of Norma Norris and Alfred Rogers. Other members of this group who made substantial contributions of their talent and work to the data analyses are Laurie Barnett, Jim Ferris, Dave Freund, Tom Jirele, Dick Harrison, Bruce Kaplan, Ed Kulick, Faustino Romero, Ira Sample, Bill Van Hassel, and Minhwei Wang. Although all of these persons made important contributions, I would be amiss not to single out the extraordinary work of Norma Norris, Al Rogers, Dave Freund, Bruce Kaplan, and Laurie Barnett.

Westat, Inc. has also contributed its wealth of wisdom and experience to this report. I am grateful to John Burke, Nancy Caldwell, Morris Hansen, Josefina Lago, Renee Slobasky, and Ben Tepping.

Making a cohesive document out of the many chapters in this report was the province of our editor, Debbie Kline, who has done an outstanding job.

Many other ETS staff members made creative contributions to the design, development, and implementation of NAEP. I am particularly indebted to Joan Baratz, Anne Campbell, Kalle Gerritz, Peg Goertz, Ann Jungeblut, and Gita Wilder.

Without an able secretarial staff, this report would have been impossible. I wish to thank Sharon Stewart, Joyce Thullen, and Marcy Yates for their hard work and dedication.

Thanks go also to Kirk Silvester and his staff, who were responsible for the printing of the final draft, and to Pete Stremic and his staff, who were responsible for the cover design and printing of the final report.

I also wish to express my gratitude to Professors Darrell Bock, John B. Carroll, Lyle Jones, and John Tukey for their important assistance. Don Searls, of the Education Commission of the States, and Jim Chromy, of the Research Triangle Institute, deserve our special thanks for their gracious assistance during the transitional period.

Albert E. Beaton
Director of Data Analysis
National Assessment of Educational
Progress
February 15, 1987

IMPLEMENTING THE NEW DESIGN:
THE NAEP 1983-84 TECHNICAL REPORT

PART I

Chapter 1

INTRODUCTION

Albert E. Beaton

Educational Testing Service

In 1982, Educational Testing Service (ETS) proposed a new, complex design for the National Assessment of Educational Progress (NAEP). The design was described extensively in The Conduct of the National Assessment of Educational Progress, a Proposal in Response to RFP PA-82-001, submitted by ETS to the National Institute of Education, November 17, 1982. An overview of the design was published in the report A New Design for A New Era (Messick, Beaton, & Lord, 1983). Three years have passed since ETS received the grant to implement its design for NAEP; the concepts in the proposed design have now been put into practice, the students have been assessed, the resulting data have been analyzed, and reports have been published. The purpose of this technical report is to report on the implementation of the 1983-84 (Year 15) assessment.

Our aim is to give the reader sufficient information to judge the utility of the design, the quality of the NAEP data, the reasonableness of the assumptions made, the appropriateness of the data analyses, and the generality of the inferences made from the data. This report covers only the technical aspects of the Year 15 NAEP. It does not attempt to provide the substantive results which might be of interest to educational policy makers; such results are provided in the reports The Reading Report Card: Progress Toward Excellence in Our Schools (1985), Writing: Trends Across the Decade, 1974-84 (Applebee, Langer, & Mullis, 1986a), and The Writing Report Card: Writing Achievement in American Schools, 1984 (Applebee, Langer, & Mullis, 1986b). The purpose of this technical documentation is to support those reports by presenting detailed information about the data and analyses that were interpreted and presented in the reports. Analyses performed specifically for these substantive reports are discussed in the procedural appendix to each report.

* * *

The National Assessment of Educational Progress is a continuing, congressionally mandated, national survey of educational achievement. The Congressional Act (Public Law 95-561-Nov. 1, 1978) under which the NAEP grant was offered states that

> [NAEP]...shall have as a primary purpose the assessment of performance of children and young adults in the basic skills of

> reading, mathematics, and communication. Such a National Assessment shall...
>
> (A) collect and report at least once every five years data assessing the performance of students at various age or grade levels in each of the areas of reading, writing, and mathematics;
>
> (B) report periodically data on changes in knowledge and skills of such students over a period of time;
>
> (C) conduct special assessments of other educational areas, as the need for additional information arises; and
>
> (D) provide technical assistance to State educational agencies and to local educational agencies on the use of the National Assessment objectives, primarily pertaining to the basic skills of reading, mathematics, and communication, and on making comparisons of such assessments with the national profile and change data developed by the National Assessment.

NAEP continues to fulfill the Congressional mandate and also gathers ancillary data which can be of use in interpreting the basic findings about the knowledge and skills of young Americans. It is the first ongoing effort to obtain comprehensive and dependable achievement data on a national basis in a uniform, scientific manner.

NAEP was originally designed and mandated by law in the 1960s, and collected its first data in 1969. The NAEP grant was administered by the Education Commission of the States (ECS) until 1983 when the grant was moved to ETS. Since its inception, NAEP has collected information not only on reading, writing, and mathematics, as required by law, but also on a number of other subject areas such as science, citizenship, art, and music. The 1983-84 (Year 15) assessment, described in this technical report, covered reading and writing as well as numerous background and attitude questions.

Before presenting the achievements of the Year 15 assessment, it is important to recall the fourteen prior years in which a National Assessment existed. During those years, the vision of Ralph Tyler, Frank Keppel, and many others was realized. As Messick, Beaton, and Lord (1983) asserted, the National Assessment design "...was brilliantly responsive to the political constraints of the time" (p. 1). The design was also brilliantly responsive to the technical constraints of the time and has been shown to have been far ahead of its time; the vision of Professor John Tukey, of Princeton University and the first Chairman of the NAEP Analysis Advisory Committee (ANAC), and many others, was indeed realized. Studying this design, and working to modify it, has brought to the ETS staff an even greater appreciation of the elegance of the original design of NAEP.

The NAEP design until 1983 included the selection of nationally representative samples of students who were 9, 13, and 17 years old, and the young adult population. Budget limitations forced the end of regular assessments of the adult population and the out-of-school 17-year-olds in the mid to late 1970s. However, these populations are assessed periodically. For efficient use of staff, the 13-year-olds were assessed in the fall of the school year, the 9-year-olds in the winter, and the 17-year-olds in the spring. Assessment exercises were assigned to students using multiple matrix sampling; different packages of exercises were assigned to students in different assessment sessions, but the same package was assigned to all students in a particular session. The assessments were administered using a tape recorder to minimize the effect of a student's reading ability on, say, his or her mathematics performance. NAEP was designed to report the achievement of students in the United States as a whole, and in subpopulations such as groups based on regions of the country, ethnicity, and gender.

The ECS design for NAEP and its modification by the ETS design are both intended to report to the interested public what students can and cannot do but differ substantially as to how to achieve that purpose. Lord (1962), who coined the term "matrix sampling", addressed the problem of estimating the proportion of a population of persons who could respond correctly to a population of items, given a fixed number of item responses. He showed that a sample with many persons responding to just one item resulted in an estimator with a smaller standard error than one derived from a sample in which fewer persons responded to many items. Of course, such sampling would not ordinarily be cost-effective, since selecting individuals is expensive, and it is usually possible to assess a number of exercises fairly inexpensively from the individuals who are sampled. The ECS conception of NAEP was interested in estimating the proportion of students who could pass particular exercises and the proportion who could pass certain pre-specified populations of exercises. Consequently, ECS' design for NAEP used a cost-effective compromise, multiple matrix sampling, which administered packages containing about 45 minutes of exercises to the students in its sample. This application of matrix sampling, however, meant that correlations could not be computed between exercises in different packages, although they could be computed between exercises within the same package. Since they were superfluous to the ECS approach, the inter-exercise correlations were seldom, if ever, used for interpreting NAEP results.

The ETS conception of NAEP is heavily dependent on knowledge of the inter-exercise correlations for expanded interpretation of the data. In simplest terms, the main idea is that if the items could be placed in such an order that a person's answering an item correctly at a particular difficulty level implied that he or she could answer all easier items, knowing the most difficult exercise a student could answer correctly would imply what that student could and could not do for the entire population of exercises. Of course, few, if any, sets of real items are so rigidly ordered, and such ordering is clearly impossible where guessing is allowed. However, other, less demanding, item response theory (IRT) models are available to be applied when the data are approximately unidimensional. Although the ECS design was sufficient to order exercises by difficulty defined in terms of percent passing an item, the inability to estimate inter-exercise correlations made it impossible to examine whether the persons who passed the more difficult exercises tended in fact to be those who passed the easier exercises, and not otherwise. BIB spiralling, a complex variant of multiple matrix sampling, was a feature of the ETS design which facilitated the collection of inter-exercise data in such a way that dimensionality could be explored.

If the dimensionality study showed that the exercises fell approximately on a single dimension, a single scale could summarize most of the information about student performance quite adequately. If the exercises fell on more than one dimension, a scale for each dimension would have to be developed, if sufficient data were available to support the scaling process; otherwise, other summarization procedures, such as the average percentages used by ECS, could be used. The 1983-84 NAEP showed that a majority of the reading exercises could be adequately fit to a unidimensional model and so these reading exercises were scaled. Using the ordering of the exercises, the reading scale was behaviorally anchored so that points on the scale could be interpreted as levels of proficiency, describing what students at those levels could and could not do. The writing exercises were scaled using an alternative method which did not require the assumption of unidimensionality.

The implementation of the ETS design for NAEP was not simple; reaching the new design goals has required some improvisation and the development of new techniques. Although ETS staff was able to begin operations about three months earlier, the NAEP grant period began on July 1, 1983 and the assessment of students began in September of the same year. First, the operational details of the old design were assimilated and merged with those of the new design. Next, the reading, writing, background, and attitude questions were reviewed and reorganized. Over 200 assessment booklets, and additional questionnaires, were printed. The cooperation of the schools was enlisted. The students were assessed and their data returned to ETS. All data was key entered, scored, and checked. Then, the data analysis began. During this time, there was continual stress between the competing goals of producing reports at the earliest possible moment and having the most carefully and elegantly constructed analysis possible. Completing a project of this magnitude and complexity required the dedication of many experts on the staffs of ETS and Westat, Inc. (the ETS subcontractor for sampling and field administration) as well as the careful coordination of their ideas and work.

The NAEP staff, of course, did not do this work alone. It had the policy guidance of the Assessment Policy Committee (APC), chaired by Wilmer Cody. It is also important to recognize the many thoughtful reviews, suggestions, comments, and other substantial help on technical issues that the NAEP staff received from the highly accomplished members of its Technical Advisory Committee (TAC), chaired by Professor Robert Linn, of the University of Illinois. Other members of the original ETS/NAEP TAC were Professor Robert Glaser of the University of Pittsburgh, Professor

Bert Green of Johns Hopkins University, Professor Sylvia Johnson of Howard University, Professor Melvin Novick of Iowa State University, and Professor Richard Snow of Stanford University. Professor Ingram Olkin of Stanford University has since replaced Professor Novick as a TAC member. The ETS staff also received important help during the transitional period from Don Searls and other members of the ECS staff and from James Chromy and others on the staff of the Research Triangle Institute (RTI).

Although this report covers all technical aspects of the Year 15 NAEP, it may be useful to summarize here the major innovative features of the Year 15 NAEP and to compare the features promised in the ETS proposal with the actuality of the Year 15 assessment.

- ETS proposed to modify the RTI sampling plan, and did. Westat, Inc. modified the sampling plan to

    (1) sample students in grades 4, 8, and 11 as well as ages 9, 13, and 17;

    (2) collect data about students whose reading and writing proficiencies could not be assessed because of physical or other handicap and who were excluded from the regular assessment sample; and

    (3) provide randomly equivalent national samples for comparing the administration procedures of the former ECS and new ETS designs.

- ETS proposed to introduce BIB spiralling, a complex method of assigning assessment exercises to students, and did. The purpose of BIB spiralling is to enhance the ability to estimate inter-exercise relationships. ETS proposed to spiral only the reading exercises but went further by spiralling together reading, writing, background, and attitude questions.

- ETS proposed to collect information on the teachers, principals, and schools of the sampled students, and did.

- ETS proposed to collect two equivalent student samples in order to measure the effect of changing from administration by tape recorder to pencil and paper, and did.

- ETS proposed to examine the dimensionality of its data in order to judge the appropriateness of scaling, and did.

- ETS proposed to scale the reading data, if appropriate, and did. The scaling procedure outlined in the ETS proposal was used, but the data were found to be too sparse at the level of individual respondents for this type of analysis. ETS then developed and applied other scaling and analytic procedures which produced satisfactory results.

- ETS proposed to form a single reading scale over all three grade/age levels, and did. The resulting NAEP reading scale spans Grade 4/Age 9 to Grade 11/Age 17 and was also used to analyze the data collected by ECS in 1970-71, 1974-75, and 1979-80.

- ETS proposed to improve the interpretability of the reading data by behaviorally anchoring various scale points, and did. A new procedure was developed to show what students at various scale points could and could not do.

- ETS did not propose to scale the writing data, but did. A new method of scaling and analysis was developed for the writing data. The writing scale was applied to all three grade/age levels assessed in 1983-84 but not to ECS' previous data since the change from a tape-recorded to a pencil-and-paper administration procedure seemed to affect writing responses substantially. As anticipated in the ETS proposal, the writing trend report was produced using the same procedures as in the past whereas the writing cross-sectional report was produced using the new writing scale.

- ETS proposed to form scales of background and attitude questions, and has done so to a small degree. A general purpose method for such scaling has been developed and applied to some writing background data, but the properties of the new method have not yet been fully explored.

- ETS proposed to run complex multivariate analyses of the NAEP data, but has not yet done so to the extent envisioned. Appropriate methods for such analyses with the NAEP data are under study. We expect more development in this area in the future.

Although the size of the grant was fixed and the actual reporting of results was not unusually slow compared to past NAEPs and other comparable surveys, the ETS design resulted in some substantial extra costs and unexpected time delays. One example of extra cost is that of collecting the inter-exercise information through BIB spiralling. This method required printing and managing over 200 different assessment booklets; about 24 booklets would have sufficed for the ECS design. The unexpected time delays resulted largely from what was essentially the research nature of this first application of the ETS design to reading and writing; when empirical results did not support the proposed analysis procedures, we developed and/or applied procedures which were more appropriate. Presumably, this research aspect of the work will be greatly reduced in future assessments of reading and writing.

* * *

The Year 15 NAEP staff was greatly concerned not only with the accuracy of its results but with making its public-use data tapes available in a format which would be as easy for others to use as possible. The purpose of the public-use data tapes is to allow others to check our analyses, to perform alternate analyses using different methods, and to perform analyses for other purposes. The public-use data tapes are already available for the Year 15 data, as they are for all previous assessments, and contain all student, teacher, and school data that were collected, except that information whose availability would risk the confidentiality of the subjects. The public-use data tapes are formatted for and have parameter statements for the commonly used statistical systems SPSS and SAS.

The dual goals of accuracy and ease of use have affected the construction of the database. Several points are worth noting.

It is impossible to make a database as complex as NAEP's completely simple to use. A secondary user cannot use the database effectively without some knowledge of the NAEP design. For example, sampling by grade and age forces the user to consider which subsample is appropriate for a particular analysis. BIB spiralling results in a substantial amount of data which is missing by design (about 90 percent!); thus, the user must think carefully about missing data procedures. Although we have tried to make the public-use data tapes as easy to use as possible, their use will require some investment in understanding NAEP.

Two features of the tapes give the user additional analytic power. Most complex surveys require sampling weights to achieve proper population estimates, and the weights are supplied for use in analysis. This has been done for NAEP. However, with a complex sampling design, the weighted versions of standard formulas for independent and identically distributed variables are not appropriate for estimating sampling errors; while appropriate formulas can be developed, they are complex to apply. Some other method based on pseudo-replicates, such as the jackknife, is appropriate and simple in application. We have developed and applied one form of the jackknife method, which we used in all NAEP analyses. It requires 32 sampling weights for each student in addition to the sampling weight usually supplied. All of these weights are available on the public-use data tapes in a way that makes possible the approximate estimation of sampling error using standard statistical systems as opposed to specialized software designed for survey data. Since this ability comes with the cost of more computing time, the secondary user may use this new ability or not, as he or she deems appropriate.

The other feature of the public-use data tapes is that they exceed the standard practice of providing only raw data by also providing derived variables for reading and writing. The complexity of the IRT scaling analysis prompted this inclusion. The underlying rationale follows.

The item-sampling designs that have characterized NAEP since its inception provided efficient estimates for average levels of performance in groups of students, but are too sparse to yield accurate estimates for individual students. Until now, NAEP reported only estimates of the proportions of students who could answer individual items or sets of items correctly, avoided estimating student proficiency distributions, and did not make individual proficiency measures available to the secondary user. The lack of individual proficiency measurements encumbered analyses of the relationships between proficiency and student characteristics. Regrettably, it is common in educational surveys to carry out these latter analyses with poorly estimated scores for individuals, despite the demonstrable invalidity of their results (see Goldstein, 1980).

Recent developments in item response theory, in statistical estimation procedures, and in methodologies for handling missing data make it possible for the first time to estimate accurately student proficiency distributions and their relationships with background variables from complex, sparse sampling designs. The embodiment of these advances, the derived variables called "plausible values" for reading and writing, were constructed to yield consistent estimates of such population characteristics for the NAEP populations as a whole, and for the subpopulations defined by the traditional NAEP reporting categories. The intricacies and expense involved in obtaining optimal estimates from such a complex database may prove prohibitive to most secondary analysts, however, and the plausible values mentioned above are therefore provided for exploratory analyses involving other background variables as well. Chapters 10 and 11 provide details on the construction and properties of plausible values and caveats on their use.

* * *

This technical report presents the details of how the assessment was accomplished, from the development of the exercises through the analyses of the data. The report is organized into three parts:

- Part I explains the steps in the process of developing the basic data. Part I begins with an overview, followed by chapters covering the development of the reading and writing exercises; the sampling; the assignment of exercises to students; a summary of the instruments; the field administration (including attainment of school cooperation); and the data entry, exercise scoring, and construction of the NAEP database and public-use data tapes. Quality control is covered throughout Part I.

- Part II explains the steps involved in data analysis. This part also begins with an overview. The next chapters include discussions of the scaling and analysis of the reading exercises, the writing exercises, and the background and attitude questions; weighting and parameter estimation, including the estimation of uncertainty due to sampling and measurement error; and the validity of the NAEP data. The final chapter of Part II discusses the use of the standard tabulations of NAEP results.

- Part III presents some estimates of the proficiency of the students in American schools. First, estimates of the numbers of students at ages 9, 13 and 17 and at grades 4, 8 and 11 are given, as well as estimates for different genders, racial/ethnic groups and other subpopulations. Then, estimates of the various points in the distributions of reading and writing proficiency are presented. Finally, estimates of average values on the reading and writing scales are given for a number of cross-classifications of students.

The organizational strategy for this report is to first present overviews of the two components of NAEP, design and analysis. These overviews direct the reader to chapters where details are provided. Each chapter begins with a summary, then presents a detailed exposition of its topic. In some cases, chapters refer to appendices or supplementary documents which contain even more detail. This strategy has been adopted to aid the reader in reaching areas of special interest. The reader who wishes only a summary may read just the overviews (Chapters 2 and 9).

We have intended to include in this report all of the avenues we have pursued, whether successfully or not, and have succeeded to some degree in doing so. This approach has been adopted to help readers understand the rationale for what was finally done and to prevent them from entering the same blind alleys in the future. Where detailed descriptions of unfruitful avenues are available, they have been included; where the wrong paths would be unduly expensive to document, they have been alluded to. We have also included some comments on what we would do differently if we could begin the design, data collection, or analysis again.

The chapters are separately authored and differ somewhat in style and point of view. In most cases, the person most responsible for the activity was assigned the writing task. We hope that the chapters can be read independently, after the appropriate introductions are read. Although we have tried to cross-reference where necessary, the method of organization results in some redundancy from chapter to chapter.

Chapter 2

OVERVIEW OF PART I: THE DESIGN AND IMPLEMENTATION OF YEAR 15 (1983-84) NAEP[1]

Albert E. Beaton

Educational Testing Service

This introduction to Part I of the technical report provides an overview of the processes by which the NAEP Year 15 data evolved from the planning stage into a database ready for analysis. The major components of this NAEP, with few details, are presented here with pointers to the succeeding chapters which contain more information. Although the remaining chapters in this part of the Technical Report contain most of the important details about each topic, some of the chapters themselves direct the reader to even greater detail to be found in appendices and supplementary documents. The organization of the report is intended to help an interested reader locate the areas of greatest interest to him or her, then study those areas in as much depth as necessary to understand the procedures and considerations involved in the collection of data. From this report, it is expected that the reader will be able to judge, for himself or herself, the quality of the data, their strengths and their weaknesses.

This chapter, and this part of the technical report, does not include a discussion of the procedures used in data analysis; the methods of data analysis are summarized in the introduction to Part II of this report, and then discussed in detail in succeeding chapters. Also, the chapter does not include the substantive results of the NAEP; those results are published separately in reports such as The Reading Report Card: Progress Toward Excellence in Our Schools (1985), etc.

Section 2.1 of this introduction provides a brief summary of the design of the Year 15 NAEP, focusing on differences between the new model for NAEP and the model which preceded it. The exposition of the design is brief; the ETS design is covered extensively in another report, A New Design for a New Era (Messick, Beaton, & Lord, 1983).

To provide background, Section 2.2 presents the NAEP assessment schedule from the first year of data collection in 1969 to the Year 15 assessment. The assessments in progress or planned through 1987-88 are also mentioned.

---

[1] The author wishes to thank Bruce Kaplan, Ira Sample, and Laurie Barnett, who produced the tables used in this chapter.

Sections 2.3 through 2.8 follow the sequence of the remaining chapters in Part I of the technical report:

- the development of the reading and writing exercises and the processes by which they were reviewed (Chapter 3);
- the four-stage stratified random sampling procedure used in the NAEP (Chapter 4);
- the assignment of the NAEP cognitive and other exercises to students selected for the sample (Chapter 5);
- a description of the instruments and an overview of the items (Chapter 6);
- the field administration procedures, including the training of the field administrators, attaining school cooperation, assessment administration, and quality control (Chapter 7); and
- the flow of data from their receipt at ETS through data entry, professional scoring, and entry into the database in final form, ready for analysis (Chapter 8).

In addition, Section 2.9 presents a statistical summary of the data that were collected in Year 15.

The data collected in the Year 15 NAEP are now ready for public use in the form of a set of public-use data tapes, documented in the NAEP 1983-84 Public-Use Data Tapes Version 3.1 Users' Guide (Barone, Norris, & Rogers, 1986). These tapes contain the available data for the sampled students, their teachers, principals, and schools.

2.1 The Design of the Year 15 Assessment

To understand the design of the Year 15 assessment, it is first helpful to review the previous design employed by the Education Commission of the States (ECS). As noted in A New Design for a New Era, the ECS design was brilliantly responsive to the demands of its times, and the ECS staff and consultants deserve substantial praise for the elegance and efficiency of that design. Because of possible variations in the definition of "grade" in different school systems, the ECS design called for sampling ages instead of grades. One of NAEP's aims was to measure performance over a broad range of exercises, while requiring not more than about 45 minutes of a student's time; thus, matrix sampling was used. To avoid the possible confounding of achievement in areas such as mathematics with the ability to read the questions and directions, tape recorders were used to present instructions and exercises. The intended result of an assessment was an estimated percent of students who could perform successfully on each exercise. The estimated percents would also be presented separately for different geographic regions, genders, races, community types, and other subgroups. The ECS staff quickly became aware that the users of their reports wanted a summarization of the massive amount of exercise-by-exercise information and thus moved to reporting, additionally, the mean percentages correct over logically homogeneous subsets of exercises and, ultimately, over all exercises within a subject area.

Nevertheless, the ETS staff felt that the original design could be modified to make NAEP results easier to understand and use, and proposed a major re-design of NAEP. ETS decided to gather samples by both age and grade because sampling only by age made the assessment results not directly relevant to school policies, which are usually established by grade level. Additionally, the tape recorders set the assessment apart from all other testing programs, so the national data from other testing programs could not be used for comparative purposes without administering those exercises using a tape recorder. The tape recorder had also resulted in the requirement that all students at an assessment session respond to the same exercises at the same moment, thus creating a less efficient sampling design. ETS proposed administration by printed instructions which would allow it to "spiral" different tests into an assessment session. ETS also introduced scaling to enhance the comparability of results over different assessment forms and with an evolving exercise pool.

ETS was sensitive to the great wealth incorporated in the data that had been collected during the previous fourteen years; data had been collected on over a million students in eleven subject areas. Any radical change which in effect made the old data unusable would not be acceptable; thus, ETS proposed to run parallel assessments in each subject area, one assessment using the past tape procedures and the other using the new printed instructions. The samples for these two assessments were equivalent; thus, differences between the two methods could be attributed to administration differences and sampling error.

2.2 Assessment Schedule

The schedule of assessments up to Year 15 is shown in Table 2(1). As this table illustrates, the subject areas assessed have included reading, writing, mathematics, science, and social studies, as well as citizenship, literature, art, music, and career development. Assessments were conducted annually through 1980 and have been conducted biennially since then. Many subject areas have been re-assessed periodically to determine trends in achievement over time. Since its inception, NAEP has assessed 9-year-olds, 13-year-olds, and in-school 17-year-olds. The assessment of out-of-school 17-year-olds and young adults was dropped because of budget restrictions. To date, NAEP has assessed approximately 1,300,000 young Americans.

Table 2(1)

NAEP Learning Areas, Grades, and Ages Assessed: 1969-1984

ASSESSMENT YEAR	LEARNING AREAS	GRADES/AGES ASSESSED*							
		Grade 4	Age 9	Grade 8	Age 13	Grade 11	Age 17IS	Age 17OS	Age ADULT
Year 1/1969-70	Science		X		X		X	X	X
	Writing		X		X		X	X	X
	Citizenship		X		X		X	X	X
Year 2/1970-71	Reading		X		X		X	X	X
	Literature		X		X		X	X	X
Year 3/1971-72	Music		X		X		X	X	X
	Social Studies		X		X		X	X	X
Year 4/1972-73	Science (2)		X		X		X	X	X
	Mathematics		X		X		X	X	X
Year 5/1973-74	Career and Occupational Development		X		X		X	X	X
	Writing (2)		X		X		X	X	
Year 6/1974-75	Reading (2)		X		X		X	X	
	Art		X		X		X	X	
Year 7/1975-76	Citizenship/ Social Studies (2)		X		X		X	X	
	Mathematics**				X		X	X	
Year 8/1976-77	Science (3)		X		X		X		
	Basic Life Skills**						X		
	Health**							X	
	Energy**							X	
	Reading** (3)							X	
	Science** (3)							X	
Year 9/1977-78	Mathematics (2)		X		X		X		
	Consumer Skills**						X		
Year 10/1978-79	Art (2)		X		X		X		
	Music (2)		X		X		X		
	Writing (3)		X		X		X		
Year 11/1979-80	Reading (4)		X		X		X	X	
	Literature (2)		X		X		X	X	
Year 12/1980-81	No Data Collection								
Year 13/1981-82	Mathematics (3)		X		X		X		
	Citizenship/ Social Studies (3)		X		X		X		
	Science** (4)		X		X		X		
Year 14/1982-83	No Data Collection								
Year 15/1983-84	Reading (5)	X	X	X	X	X	X		
	Writing (4)	X	X	X	X	X	X		

* 17IS denotes 17-year-olds enrolled in public or private schools; 17OS denotes 17-year-olds who dropped out of school or graduated prior to the assessment.

** Small, special-interest assessments conducted on limited samples at specific ages

( ) Second and subsequent assessments of a learning area

1983-84 was a transition year for NAEP. The Education Commission of the States (ECS) had the role of deciding the subject areas to be measured, reading and writing, and developing the exercises and background and attitude items to be administered. The Research Triangle Institute (RTI), the sampling and field administration subcontractor for ECS, had the role of selecting the sample of schools. The Educational Testing Service (ETS) prepared the assessment booklets according to its design, prepared the data for analysis, and analyzed the data. Westat, Inc., the sampling and field administration subcontractor for ETS, modified and extended the sample and administered the assessment to the sampled students.

Table 2(1) also indicates the initiation in Year 15 of data collection by grade as well as by age.

Assessments through 1988 are either in progress or in the planning stage. In Year 16 (1984-85), a separately funded assessment of the literacy of young adults was administered, the results of which have been published in Literacy: Profiles of America's Young Adults, Final Report (Kirsch & Jungeblut, 1986) This survey also collected a small sample of out-of-school 17-year-olds. The Year 17 (1985-86) assessment includes reading, mathematics, science, and computer competence, with a special probe of U.S. history and literature for the older students. Current plans call for the assessment of reading, writing, citizenship, and U.S. history in Year 19 (1987-88).

2.3 The Development of NAEP Measurement Instruments

The Year 15 NAEP assessed the performance of American students in the learning areas of reading and writing. In addition, a large number of background and attitude questions were surveyed. Information was also collected from the students' principals and teachers.

The development of the reading exercises, writing exercises, and background and attitude items was the responsibility of ECS. ECS gave to ETS a large number of exercises, more than could be used, and ETS selected the items that were actually administered. The details of the development of the exercises are provided in Chapter 3.

From its inception, NAEP has developed assessments through a consensus process. Educators, scholars, and citizens representative of many diverse constituencies and points of view design objectives for each subject area assessment, proposing general goals they feel students should achieve in the course of their education. After careful reviews, the objectives are given to item writers, who develop assessment questions appropriate to the objectives.

All exercises undergo extensive reviews by subject-matter and measurement specialists, as well as careful scrutiny to eliminate any potential bias or lack of sensitivity to particular groups. Some of the questions used in each assessment are made available to anyone interested in studying or using them. The remainder have traditionally been kept secure for use in future assessments for the examination of trends over time.

The reading assessment contained multiple choice, short open-ended, and essay exercises. The reading essay exercises were professionally scored. All writing exercises required that the student write an essay, and these essays were also professionally scored. The professional scoring is described in Chapter 8.2.

In recent assessments, NAEP has asked numerous background and attitude questions to improve the usefulness of NAEP achievement results and provide the opportunity to examine policy issues. Students, teachers and school officials answer a variety of questions about instruction, activities, experiences, curricula, resources, attitudes and demographics.

2.4 The NAEP Sample

The sampled populations consisted of all in-school students who were 9, 13, or 17 years old or who were in the 4th, 8th, or 11th grades in the 50 states and the District of Columbia. Both public and private school students were sampled.

The sample is a four-stage probability sample. The stratified sample of first-stage units and of schools within selected first-stage units was developed and selected by the Research Triangle Institute, using sample sizes specified by Westat, Inc. The third and fourth stages of sampling, involving the assignment of sessions to schools and the selection of students were designed and implemented by Westat, Inc. The Year 15 sample design is described in detail in Chapter 3 and in Westat's Report on Sample Selection, Weighting, and Variance Estimation: NAEP--Year 15 (Lago, Burke, Tepping, & Hansen, 1985). The four stages are described below. The details of the sampling procedure can be found in Chapter 4.

Stage 1: Primary Sampling Units

In the first stage of sampling, the United States was divided into geographical units comprised of counties or groups of contiguous counties, which met a minimum school enrollment size. These units, called primary sampling units (PSUs), were classified into 20 strata which were defined by region (Northeast, Southeast, Central, and West) and by the sample description of community (Big City, Fringe of Big City, Medium City, Small Place, and Extreme Rural). A sample of 64 PSUs was then selected (without replacement) to represent all regions and sizes of communities with probability proportional to population size measures.

Stage 2: Sampling Schools

In the second stage of sampling, the frame consisted of a file of schools obtained from Quality Education Data, Inc. (QED). The file included public, private, Catholic, Bureau of Indian Affairs, and Department of Defense schools, listed according to the three grade/age groups within each of the 64 PSUs. The NAEP grade/age groups were Grade 4/Age 9, Grade 8/Age 13, and Grade 11/Age 17.

To allow sampling of extreme-low socio-economic status (SES) big-city schools at a double sampling rate, schools within big-city PSUs were stratified by SES and their estimated sizes were doubled. Extreme-rural schools were also oversampled by a factor of two.

Schools within each PSU were selected (without replacement) with probabilities proportional to assigned measures of size. Roughly equal measures of size were assigned to schools containing estimates of age-eligible students ranging from 20 to 160 (for age 9), or 20 to 200 (for ages 13 and 17). Schools above the indicated maximum size were selected with probabilities proportional to the number of age-eligible students. Schools with less than 20 estimated age eligibles were assigned considerably lower measures of size, since they had higher per-student administrative costs.

Stage 3: Assignment of Sessions to Schools, by Type

The assignment of sessions to schools served as the third stage of sampling. This assignment was done separately by the two types of sessions, designated "spiral" and "tape", which represent separate samples of the population of students. The Year 15 tape sample contains students of specified ages (who could be of any grade). The Year 15 spiral sample contains the students who received either BIB or UBIB booklets (see below) and represents two overlapping samples. The first sample represents students of specified ages (who could be in any grade) and is comparable to the samples from previous NAEP assessments. It is also randomly equivalent to the samples of students who were administered the tape booklets in the Year 15 assessment. The second sample represents students of specified grades (who could be of any age).

For tape assessments there were four distinct booklets at each age class, each of which was to be administered once within each of the PSUs. To assure that no tape session would include a very small number of students, small schools were clustered with other schools in the same PSU so as to form clusters with an estimated minimum of eight eligibles. Tape sessions were then assigned within each PSU by ordering schools or school clusters by socioeconomic status and size and then selecting a systematic sample of four schools at each age with probability proportional to the estimated number of age eligibles within the school (or school cluster).

One spiral session was assigned to each school or school cluster which was not selected for a tape session. The balance of the spiral sessions were then assigned to schools (and school clusters) at a rate approximately proportional to the estimated number of eligible students, by age or grade, that would be available after the initial assignment of tape and spiral sessions.

Stage 4: Sampling Students

In the fourth stage of sampling, a consolidated list of all grade- and age-eligible students was established for each selected school. A systematic selection of eligible students was made and students were assigned to spiral or tape sessions, depending on whether the assessment was to be administered by pencil and paper or by tape recorder.

Some students were deemed unassessable by the school authorities because they did not speak English, were judged as being educable mentally retarded, or were functionally disabled. In these cases, a questionnaire was filled out by the school staff listing the reason for excluding the student and providing some background information.

Sampling Principals and Teachers

A Principal Questionnaire, distributed to each sampled school by Westat prior to the assessment, was used by Westat to obtain both an up-to-date estimate of grade/age-eligible students and information on minority enrollment.

The School Characteristics and Policy Questionnaire and Teacher Questionnaire were distributed in every sampled school. The School Characteristics and Policy Questionnaire was mailed to the school by Westat prior to the assessment and picked up by the Westat supervisor, then returned to ETS.

The Teacher Questionnaire was administered to the teachers of a subsample of the students sampled for spiral sessions. The purpose of this sample was to estimate the number (proportion) of students whose teachers had various attributes, not the attributes of the teacher population. Therefore, statements like "20 percent of students have teachers who have..." are appropriate in discussing Teacher Questionnaire data, but statements like "20 percent of teachers have..." are not.

The number of teachers sampled was equal to the number of spiral sessions conducted in the school. Thus, if there were six spiral sessions conducted in a school, a subsample of six students was selected and the school coordinator was asked to identify the English or Language Arts instructor for each student. These instructors completed the Teacher Questionnaires. Please note that, since a number of students may have had the same teacher, and some teachers did not complete the questionnaire, the number of students in the subsample for whom teacher information is available is greater than the number of teachers who completed questionnaires in a given school.

2.5 Assigning NAEP Exercises to Students

After the student sample was selected, it was necessary to assign booklets of exercises to them. The ETS design for NAEP greatly affected the way in which the assessment booklets were organized and constructed. The assignment of booklets depended on whether the student was in the spiral or the tape samples. A detailed discussion of this topic can be found in Chapter 5.

2.5.1 Spiral Booklets

The spiral sample is so called because the assessment booklets were spiralled within an assessment session, that is, different booklets were interleaved so that different students in the same assessment session were asked to respond to different exercises. With spiralling, the instructions to the students and the exercises themselves must be read by the student from his or her booklet since administration using a tape recorder would be unmanageable with more than one type of booklet in an assessment session. The purpose of spiralling was to increase sampling efficiency.

The targeted sample size was for 2,000 students to respond to each exercise at each age or grade level in the spiral sample; this target implied a sample of 2,600 at each grade/age.

The reading and writing exercises were sorted into units called blocks which were designed to take a student fourteen minutes to complete. The fourteen minutes included, on the average, twelve minutes of either reading or writing exercises and two minutes of background and attitude questions. Altogether, there were 21 such blocks of exercises created for each grade/age level. Three double-length (28-minute) blocks were also developed, making 24 blocks per grade/age combination. Some blocks were administered at more than one age and grade.

The spiral sample can be divided into two parts: the BIB and UBIB samples.

The BIB (Balanced Incomplete Block) sample was created to meet the design goal of facilitating the estimation of inter-correlations or other statistics among the assessment exercises. Using a BIB design, a large number of booklets were created in such a way that each pair of exercises was administered to a randomly equivalent subsample of students while maintaining the goal of 2,000 students for each exercise at both age and grade levels. For the BIB part of the spiral sample, 57 assessment booklets were assembled for each grade/age level. Each booklet began with one six-minute block which contained only background questions which was followed by three fourteen-minute blocks containing a combination of cognitive exercises and background and attitude questions.

The UBIB (Unbalanced Incomplete Block) part of the spiral sample was developed to accommodate several long exercises which could not fit into the 14-minute blocks, thus necessitating the development of three double-length blocks. Because of these three blocks, six more booklets, called UBIB booklets, were created using an unbalanced design. The UBIB booklets began with the same common background section, which was followed by one double-length and one single-length block. Since it was not possible to pair the double-length blocks with each other within the available student time, some of the single-length blocks used in the UBIB booklets were also used in the BIB booklets and thus are available for selected inter-correlations.

The booklets developed using the BIB and UBIB designs were interleaved into the spiral sample in bundles of 23 in a randomly selected order. Each of the 57 BIB and 6 UBIB booklets were placed in the bundles in such a way that the estimated number of students receiving each block was at least 2,600 per grade/age. Each booklet had the appropriate probability of being at each position within a bundle.

The bundles were distributed in assessment sessions within a school so that, in almost all instances, no two students in a session were given the same booklet.

2.5.2 Tape Booklets

Another design feature of the new NAEP was the collection of bridge samples by administering some of the NAEP items with tape recordings and age-only simple matrix sampling as had been done in past assessments. The purpose of these samples was to explore the effect of the change from tape-recorded administration to pencil-and-paper administration and, if possible, to project the results of past assessments onto the new scale (see Chapters 10 and 11). Using some of the items in the BIB and UBIB booklets, four tape booklets were administered to age-eligible students of each grade/age.

Thus, another four booklets were printed for each grade/age level for administration using NAEP's former procedures. These booklets were administered in separate sessions using a tape recorder for directions. All students in a given tape session were administered the same booklet. Each of the four tape booklets contained two sections: a section of common background items and a section of cognitive reading and writing exercises (three booklets contained both reading and writing exercises in the cognitive section; one contained only reading exercises). In most cases, the cognitive exercises were also used in past assessments and in the spiral booklets.

2.5.3 Timing

The Grade 8/Age 13 students were assessed in the fall of 1983, the Grade 4/Age 9 students were assessed in the winter of 1984, and the Grade 11/Age 17 students were assessed in the spring of 1984.

A testing session lasted approximately one hour. The BIB and UBIB booklets took 48 minutes of actual testing time; tape booklets took approximately 53 minutes.

See Chapter 5 for details.

2.6 Instrument and Item Information

The assessment incorporated four distinct types of instruments: student assessment booklets; a questionnaire for excluded students; a teacher questionnaire; and a school characteristics and policies questionnaire.

The student assessment booklets were composed of items that were either cognitive or non-cognitive. Cognitive items were reading exercises, study skill exercises or writing exercises. Non-cognitive items asked questions relative to the backgrounds and attitudes of students. Some non-cognitive items were presented to every student and were placed together in a block called the common block or common core. Others were placed at the beginning of the blocks containing the cognitive items.

The reading items included short and long reading passages, graphically presented materials, poems, and reference materials (e.g., tables of contents). Some items required a multiple-choice response, some open-ended items required a brief written response, and some required written essays. A total of 176 reading items was presented to Grade 4/Age 9; a total of 192 reading items was presented to Grade 8/Age 13; and a total of 196 reading items was presented to Grade 11/Age 17.

The writing items were developed to assess performance in three writing areas: informative, persuasive and imaginative. Students were asked to write, for example, letters, descriptive essays, or narrative pieces. From a total pool of 22 writing items, 15 were used at each grade/age.

Each booklet included six minutes of background and attitude items common to all students. These items are general questions concerning materials in the home, parental education, etc. Each block also contained additional background and attitude items, related to objectives formulated for reading and writing. The items measured students' perceptions of their teachers' instructional practices in reading and writing; their own study habits and reading activities; their perceptions of the value of reading and writing; and their assessment of themselves as readers and writers.

The Excluded Student Questionnaire was developed and used for the first time in the Year 15 assessment. It was designed to gather more information about particular conditions for exclusion and characteristics of the learning experience of excluded students.

The Teacher Questionnaire was also developed and used for the first time in Year 15. It was designed to gather information on the curricula and teaching methods used by selected English and Language Arts teachers.

The School Characteristics and Policy Questionnaire was distributed to each participating school to be completed by either the school's principal or another person familiar with data concerning enrollment, facilities, curricula and staff development.

More information about the items and instruments can be found in Chapter 6.

2.7 Field Administration

Westat was responsible for field administration. The process began with the development of necessary materials and a field organization. Materials were developed for training, contacting the schools, sampling, and process control. The field organization consisted of district supervisors and exercise administrators. Westat trained the district supervisors, who in turn trained the exercise administrators.

Gaining school cooperation was a joint effort of Westat and ETS. ETS first contacted the Chief State School Officers (CSSOs), informing them that schools within their states had been selected for NAEP. Later, mailings and materials were sent to the CSSOs, school district superintendents and private school officials. Meeting arrangements were then established by telephone and contact forms were filed with Westat. Westat district supervisors then scheduled and conducted introductory meetings.

Westat administered the assessment in the field primarily through the work of district supervisors. District supervisors had many responsibilities, including drawing the sample of students, completing assessment reporting forms, making final arrangements for the assessments, supervising exercise administrators, distributing and collecting other data forms and questionnaires, and editing, boxing and shipping assessment materials.

Both Westat and ETS were responsible for quality control. There were two specifically designed quality control studies of the field effort. The first, and most intensive, involved on-site visits by Westat and ETS staff to verify the sampling and to observe the supervisors and exercise administrators as they conducted assessments. The second study was a telephone survey of a ten-percent sample of schools. This survey took place after the field period had ended and all assessment activities had been completed in the schools.

Field administration is discussed in detail in Chapter 7.

2.8 Database Construction

Westat shipped the assessment booklets to ETS for entry into computer files, checking, and forming the database. Careful checking assured that all data from the field were received. The data then went through extensive processing, described in Chapter 8.

Since machine readable assessment booklets were not used, an "intelligent" data entry system was developed. This computer program not only received the input data but also checked for consistency among the many different booklets, blocks, and formats. The program assured that all entered values of each variable were within the range of possible values. All data were independently key-verified and all discrepancies were resolved.

Student responses to some of the reading exercises and all of the writing exercises had to be professionally scored. Professional scorers were hired, trained, and closely supervised. Exercises were scored by both holistic and primary trait methods, as well as some secondary trait and mechanics scoring methods. Random samples of essays were independently re-scored and reliability coefficients were estimated.

Extensive quality control checks were instituted to assure correspondence between what had been written in the booklet and what appeared in the database. A random sample of each assessment booklet and questionnaire was selected from the computer file and checked against the original document. The database was determined to be extraordinarily error-free.

The construction of the database and public-use data tapes are described in more detail in Chapter 8.

2.9 Tabular Summary of NAEP Year 15 Sample

The purpose of this section is to present the characteristics of the Year 15 (1983-84) NAEP data in a tabular form. This section is a statistical summary of the results of the data collection steps outlined above and is intended to describe the sample, not to estimate the characteristics of the population of American students.

There were three samples of students which were defined by being at either a particular age or a particular grade level:

Age Class 1: Grade 4/Age 9
Age Class 2: Grade 8/Age 13
Age Class 3: Grade 11/Age 17

This sample was designed for estimating population values defined either by age or by grade; for example, the sample of age 9 students includes 9-year-old students in other grades as well as the fourth grade, and the grade 4 sample contains fourth graders of all ages, not just 9-year-olds.

The system of defining age that was used in past assessments was maintained in Year 15; thus, the dates of birth were defined as follows:

Age 9: Born between January 1 and December 30, 1974
Age 13: Born between January 1 and December 30, 1970
Age 17: Born between October 1, 1966 and September 30, 1967

A student's grade level was defined by the school. Note that only 17-year-olds who were enrolled in school were sampled in Year 15; out-of-school 17-year-olds were not sampled.

2.9.1 Measurement Instruments

The measurement instruments that were produced by ETS are summarized in Table 2(2). The same number of instruments were produced at each grade/age level. In addition to these instruments, some school level data were available from a Principal Questionnaire developed by Westat and from the QED database that was used in developing the sampling frame.

The number of items used in the measurement instruments varies from one age class to another. The item counts are shown in Table 2(3). The Total column is not the sum of the three grade/age columns because some items were used for more than one age class.

The reading and writing exercises were placed in blocks, and the blocks placed in booklets that were administered to either the spiral sample, the tape sample, or both. The assignment of exercises to types of administration is shown in Table 2(4).

2.9.2 PSU, School, and Teacher Sample Characteristics

Table 2(5) shows the distribution of Primary Sampling Units (PSUs) that were selected for the Year 15 sample by region of the country and by the sampling description of community. The sampling frame called for 20 cells (4 regions by 5 types of community) but the Northeast had so few small places and extreme rural counties (and pseudo-counties) that they were combined for sampling purposes. These combined strata are shown as a separate column in this table. The same PSUs were used for all age classes.

The cooperation rates and the characteristics of the schools participating in this NAEP are shown in Table 2(6). There was a total of 1,480 schools in the assessment of which 1,465 have data for at least one student. A total of 1,382 schools returned the school questionnaire. We have used the same method of computing cooperation rates as used in the Year 13 assessment, and the cooperation rates were taken from the report, Year 13 Field Operations and Data Collection Activities (Research Triangle Institute, 1982). The Year 15 figures were taken from Westat's Report on Sample Selection, Weighting, and Variance Estimation: NAEP--Year 15 (Lago, Burke, Tepping, & Hansen, 1985). This table also shows the distribution of schools by region, school affiliation, size and type of community, urbanicity, grade span, number of teachers, and number of students.

The count of teacher questionnaires is shown in Table 2(7). A questionnaire was returned by a total of 2,732 teachers over all age classes. Of these teachers, 2,685 taught at least one student for whom data are available. A total of 52,367 students could be associated with a teacher who returned a questionnaire.

The number of assessment sessions, including makeup sessions, is shown in Table 2(8). The sessions are shown separately by spiral and tape administrations.

2.9.3 Student Sample Characteristics

Data were collected for a total of 104,437 students in Year 15. The number of students who were administered the spiral assessment and tape assessment are shown in Table 2(9). This table also includes the number of students who were deemed by the school to be unable to respond to the assessment situation and were thus excluded from the sample.

Tables 2(10), 2(11), and 2(12) show the sizes of various subsamples from the spiral sample of students for the different grade/age levels. Subsamples are defined by sex, race, region of the country, parents' education, and size and type of community. Sample sizes are shown separately for age eligibles, grade eligibles, age and grade eligibles, and for the entire age class.

Tables 2(13), 2(14), and 2(15) show the sizes of the same subsamples for the excluded students by grade/age level. Sample sizes are shown separately for age eligibles, grade eligibles, and age and grade eligibles as well as for the entire age class.

Tables 2(16), 2(17), and 2(18) show the sizes of the same subsamples for the four tape-administered instruments. These samples are age eligibles only.

* * *

The Year 15 data are now available on a set of public-use data tapes, for which the accompanying NAEP 1983-84 Public-Use Data Tapes Version 3.1 Users' Guide (Barone, Norris, & Rogers, 1986) has been prepared. The public-use data tapes contain all student, teacher, and school data except information that was excised to preserve the respondents' anonymity. Because the sampled students did not have an equal probability of selection, the sampling weights are included on the data tapes. The current edition of the public-use data tapes also contains some derived variables such as reading and writing proficiency estimates. The public-use data tapes developed from previous assessments by the Education Commission of the States are also still available.

Table 2(2)

Measurement Instruments
Developed by ETS

	Grade/Age		
	4/9	8/13	11/17
BIB BOOKLETS	57	57	57
UBIB BOOKLETS	6	6	6
TAPE BOOKLETS	4	4	4
EXCLUDED STUDENT QUESTIONNAIRES	1	1	1
TEACHER QUESTIONNAIRES	1	1	1
SCHOOL CHARACTERISTICS QUESTIONNAIRES	1	1	1
TOTAL	70	70	70

Table 2(3)

Number of Items Administered

	Grade/Age			
	4/9	8/13	11/17	TOTAL
READING	176	192	196	340
WRITING	15	15	15	22
BACKGROUND AND ATTITUDE	260	273	376	378
EXCLUDED STUDENTS	72	72	72	72
TEACHER	351	293	293	417
SCHOOL CHARACTERISTICS	247	242	247	321
TOTAL	1121	1087	1199	1550

Table 2(4)

Number of Reading and Writing Exercises by Type of Administration

	Grade/Age		
	4/9	8/13	11/17
READING:			
SPIRAL ONLY	78	88	92
TAPE ONLY	1	0	20
SPIRAL AND TAPE	97	104	84
TOTAL	176	192	196
WRITING:			
SPIRAL ONLY	12	12	12
TAPE ONLY	0	0	0
SPIRAL AND TAPE	3	3	3
TOTAL	15	15	15

Table 2(5)

Allocation of PSUs to Regions and Community Types

	Big City	Urban Fringe	Medium City	Small Places	Extreme Rural	Northeast Small Places & Extreme Rural	Total
NORTHEAST	5	3	4	-	-	2	14
SOUTHEAST	3	2	4	5	2	-	16
CENTRAL	5	3	4	3	3	-	18
WEST	8	1	3	2	2	-	16
TOTAL	21	9	15	10	7	2	64

Table 2(6)

Characteristics of Schools

	Grade/Age			
	4/9	8/13	11/17	TOTAL
TOTAL NUMBER OF SCHOOLS	663	486	331	1480
NUMBER WITH DATA*	661	478	326	1465
NUMBER WITH COMPLETED QUESTIONNAIRES	623	457	302	1382
COOPERATION RATE:				
YEAR 15	88.6	90.3	83.9	88.1
YEAR 13	88.0	89.2	86.5	88.0
REGION:				
NORTHEAST	151	99	69	319
SOUTHEAST	145	116	80	341
CENTRAL	222	162	99	483
WEST	145	109	83	337
SCHOOL TYPE:				
PUBLIC	522	337	281	1140
PRIVATE	42	46	31	119
CATHOLIC	97	102	19	218
NO INFORMATION	2	1	0	3
SIZE AND TYPE OF COMMUNITY:				
EXTREME RURAL	69	54	36	159
LOW METROPOLITAN	68	51	31	150
HIGH METROPOLITAN	69	49	37	552
MAIN BIG CITY	56	47	25	128
URBAN FRINGE	58	45	23	126
MEDIUM CITY	93	62	46	201
SMALL PLACE	250	178	133	561
URBANICITY:				
URBAN	211	183	82	476
SUBURBAN	207	120	110	437
RURAL	243	182	139	564
NO INFORMATION AVAILABLE	2	1	0	3

* Several schools were sampled but no eligible students were selected. These schools are retained in the NAEP database.

Table 2(6)

Characteristics of Schools
(continued)

	Grade/Age			
	4/9	8/13	11/17	TOTAL
GRADE SPAN:				
KINDERGARTEN TO GRADE 12	24	32	30	86
KINDERGARTEN TO GRADE 3	19	0	0	19
KINDERGARTEN TO GRADE 6	427	28	0	455
KINDERGARTEN TO GRADE 8	175	204	1	380
GRADE 6 OR 7 TO GRADE 8	16	114	0	130
GRADE 7 TO GRADE 9	0	41	7	48
GRADE 7 TO GRADE 12	0	58	58	116
GRADE 9 TO GRADE 12	0	8	194	202
GRADE 10 TO GRADE 12	0	0	41	41
NO INFORMATION	2	1	0	3
NUMBER OF TEACHERS:				
1 - 4	22	12	3	37
5 - 9	112	58	9	179
10 - 19	251	144	50	445
20 - 49	265	209	113	587
50 - 74	9	45	56	110
75 - 99	1	10	55	66
100+	1	6	44	51
NO INFORMATION	2	2	1	5
NUMBER OF STUDENTS:				
1 - 99	30	15	7	52
100 - 299	245	134	50	429
300 - 499	240	125	61	426
500 - 749	116	116	43	275
750 - 999	20	48	27	95
1000 - 1499	10	34	55	99
1500+	0	13	87	100
NO INFORMATION	2	1	1	4

Table 2(7)

Number of Responses to Teacher Questionnaire

	Grade/Age			
	4/9	8/13	11/17	TOTAL
TEACHERS*	1027	790	915	2732
TEACHERS WITH STUDENTS IN SAMPLE	1005	779	901	2685
STUDENTS WITH TEACHERS**	14846	20838	16673	52357

* Some teachers responded but were not linked to any student in the sample.
** Teachers were often associated with many students.

Table 2(8)

Number of Assessment Sessions by Type of Administration

	Grade/Age			
	4/9	8/13	11/17	TOTAL
NUMBER OF SESSIONS:				
SPIRAL REGULAR	1328	1327	1327	3982
SPIRAL MAKEUP*	2	4	93	99
TOTAL	1330	1331	1420	4081
NUMBER OF SESSIONS:				
TAPE REGULAR	259	258	256	773
TAPE MAKEUP*	0	0	67	67
TOTAL	259	258	323	840

* See Section 7.3.2 for details about makeup sessions.

Table 2(9)

Number of Students
by Type of Administration

	Grade/Age			
	4/9	8/13	11/17	TOTAL
SPIRAL	26087	28405	28861	83353
TAPE	5492	5158	6209	16859
EXCLUDED	1416	1448	1361	4225
TOTAL	32995	35011	36431	104437

Table 2(10)

Spiral Sample by Demographic Characteristics

Grade 4/Age 9

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL	18945	20095	12953	26087
SEX:				
MALE	9496	10213	6091	13618
FEMALE	9449	9882	6862	12469
RACE:				
WHITE	12635	13272	8920	16987
BLACK	2800	3162	1819	4143
HISPANIC	2640	2777	1614	3803
OTHER	870	884	600	1154
REGION:				
NORTHEAST	4257	4579	3227	5609
SOUTHEAST	4744	5110	3198	6656
CENTRAL	5380	5544	3547	7377
WEST	4564	4862	2981	6445
PARENTS ED:				
LESS THAN HIGH SCHOOL	1089	1285	670	1704
HIGH SCHOOL	3628	4106	2524	5210
GREATER THAN HIGH SCHOOL	6885	7465	5086	9264
UNKNOWN	7343	7239	4673	9909
STOC:				
RURAL	1204	1305	760	1749
DISADVANTAGED URBAN	2490	2721	1698	3513
ADVANTAGED URBAN	2216	2336	1724	2828
BIG CITY	1500	1503	1060	1943
FRINGE	1991	2068	1423	2636
MEDIUM	2913	3097	1915	4095
SMALL	6631	7065	4373	9323

Table 2(11)

Spiral Sample by Demographic Characteristics

Grade 8/Age 13

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL	21070	21850	14515	28405
SEX:				
MALE	10526	10928	6774	14680
FEMALE	10543	10920	7740	13723
RACE:				
WHITE	15047	15525	10820	19752
BLACK	2922	3099	1774	4247
HISPANIC	2398	2471	1428	3441
OTHER	703	755	493	965
REGION:				
NORTHEAST	4730	4956	3608	6078
SOUTHEAST	5191	5514	3571	7134
CENTRAL	6041	6119	4049	8111
WEST	5108	5261	3287	7082
PARENTS ED:				
LESS THAN HIGH SCHOOL	1870	2185	1112	2943
HIGH SCHOOL	7427	7751	5079	10099
GREATER THAN HIGH SCHOOL	9495	9753	7105	12143
UNKNOWN	2278	2161	1219	3220
STOC:				
RURAL	1215	1308	796	1727
DISADVANTAGED URBAN	2189	2188	1369	3008
ADVANTAGED URBAN	2319	2387	1867	2839
BIG CITY	2225	2221	1595	2851
FRINGE	2841	2977	2097	3721
MEDIUM	2769	2981	1842	3908
SMALL	7512	7788	4949	10351

Table 2(12)

Spiral Sample by Demographic Characteristics

Grade 11/Age 17

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL	22783	22865	16787	28861
SEX:				
MALE	11327	11294	8006	14615
FEMALE	11454	11571	8781	14244
RACE:				
WHITE	16482	16681	13017	20146
BLACK	3345	3331	1986	4690
HISPANIC	2192	2054	1285	2961
OTHER	7640	799	499	1064
REGION:				
NORTHEAST	5097	5185	3772	6510
SOUTHEAST	5766	5817	4016	7567
CENTRAL	6391	6355	5004	7742
WEST	5529	5508	3995	7042
PARENTS ED:				
LESS THAN HIGH SCHOOL	2806	2761	1740	3827
HIGH SCHOOL	8018	7883	5866	10035
GREATER THAN HIGH SCHOOL	10957	11277	8596	13638
UNKNOWN	1002	944	585	1361
STOC:				
RURAL	1381	1473	1063	1791
DISADVANTAGED URBAN	2461	2329	1381	3409
ADVANTAGED URBAN	2968	3060	2280	3748
BIG CITY	2161	2139	1487	2813
FRINGE	2272	2223	1706	2789
MEDIUM	3804	3848	2939	4713
SMALL	7736	7793	5931	9598

Table 2(13)

Excluded Student Sample by Demographic Characteristics

Grade 4/Age 9

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL*	966	826	376	1416
SEX:				
MALE	614	532	230	916
FEMALE	351	291	145	497
RACE:				
WHITE	460	369	160	669
BLACK	125	121	52	194
HISPANIC	275	247	123	399
OTHER	106	89	41	154
REGION:				
NORTHEAST	229	190	105	314
SOUTHEAST	224	167	68	323
CENTRAL	212	197	81	328
WEST	301	272	122	451
STOC:				
RURAL	47	53	13	87
DISADVANTAGED URBAN	238	212	94	356
ADVANTAGED URBAN	94	69	49	114
BIG CITY	78	57	37	98
FRINGE	113	102	60	155
MEDIUM	130	112	35	207
SMALL	266	221	88	399

*Some demographic subgroups do not add up to Total due to missing or unresolved subgroup data.

Table 2(14)

Excluded Student Sample by Demographic Characteristics

Grade 8/Age 13

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL*	907	901	360	1448
SEX:				
MALE	579	561	219	921
FEMALE	322	338	140	520
RACE:				
WHITE	430	474	175	729
BLACK	202	172	68	306
HISPANIC	182	172	78	276
OTHER	93	83	39	137
REGION:				
NORTHEAST	154	159	69	244
SOUTHEAST	216	243	74	385
CENTRAL	290	291	119	462
WEST	247	208	98	357
STOC:				
RURAL	34	49	11	72
DISADVANTAGED URBAN	204	183	95	292
ADVANTAGED URBAN	54	52	14	92
BIG CITY	108	98	45	161
FRINGE	88	73	36	125
MEDIUM	105	124	29	200
SMALL	314	322	130	506

*Some demographic subgroups do not add up to Total due to missing or unresolved subgroup data.

Table 2(15)

Excluded Student Sample by Demographic Characteristics

Grade 11/Age 17

	AGE ELIGIBLE	GRADE ELIGIBLE	AGE & GRADE ELIGIBLE	TOTAL
TOTAL*	983	707	329	1361
SEX:				
MALE	640	458	218	880
FEMALE	342	248	111	479
RACE:				
WHITE	371	311	135	547
BLACK	237	141	60	318
HISPANIC	249	168	98	319
OTHER	126	87	36	177
REGION:				
NORTHEAST	130	94	40	184
SOUTHEAST	258	176	79	355
CENTRAL	267	218	84	401
WEST	328	219	126	421
STOC:				
RURAL	48	38	20	66
DISADVANTAGED URBAN	285	154	89	350
ADVANTAGED URBAN	59	54	26	87
BIG CITY	106	74	31	149
FRINGE	74	71	28	117
MEDIUM	163	122	54	231
SMALL	248	193	81	360

*Some demographic subgroups do not add up to Total due to missing or unresolved subgroup data.

Table 2(16)

Tape Sample by Demographic Characteristics

Grade 4/Age 9

	TAPE 1	TAPE 2	TAPE 3	TAPE 4	TOTAL
TOTAL	1403	1356	1389	1344	5492
SEX:					
MALE	691	701	696	653	2741
FEMALE	712	655	693	691	2751
RACE:					
WHITE	970	869	914	832	3585
BLACK	182	223	246	178	829
HISPANIC	186	203	181	263	833
OTHER	65	61	48	71	245
REGION:					
NORTHEAST	310	335	275	288	1208
SOUTHEAST	348	339	349	347	1383
CENTRAL	409	365	411	364	1549
WEST	336	317	354	345	1352
PARENTS ED:					
LESS THAN HIGH SCHOOL	81	76	83	104	344
HIGH SCHOOL	247	280	284	277	1088
GREATER THAN HIGH SCHOOL	567	472	509	453	2001
UNKNOWN	508	528	513	510	2059
STOC:					
RURAL	148	114	128	74	464
DISADVANTAGED URBAN	156	194	254	205	809
ADVANTAGED URBAN	260	183	178	90	711
BIG CITY	130	62	168	199	559
FRINGE	117	152	49	70	388
MEDIUM	243	136	136	101	616
SMALL	349	515	476	605	1945

Table 2(17)

Tape Sample by Demographic Characteristics

Grade 8/Age 13

	TAPE 1	TAPE 2	TAPE 3	TAPE 4	TOTAL
TOTAL	1310	1276	1283	1289	5158
SEX:					
MALE	676	636	637	680	2629
FEMALE	634	639	646	609	2528
RACE:					
WHITE	945	889	844	915	3593
BLACK	187	211	226	160	784
HISPANIC	121	126	165	178	590
OTHER	57	50	48	36	191
REGION:					
NORTHEAST	275	286	262	259	1082
SOUTHEAST	327	334	329	333	1323
CENTRAL	388	356	366	361	1471
WEST	320	300	326	336	1282
PARENTS ED:					
LESS THAN HIGH SCHOOL	129	92	140	114	475
HIGH SCHOOL	464	451	487	470	1872
GREATER THAN HIGH SCHOOL	600	574	527	567	2268
UNKNOWN	117	159	129	138	543
STOC:					
RURAL	126	62	73	146	407
DISADVANTAGED URBAN	162	141	206	113	622
ADVANTAGED URBAN	12	81	107	123	437
BIG CITY	142	102	75	126	445
FRINGE	194	240	213	180	827
MEDIUM	184	209	168	224	785
SMALL	376	441	441	377	1635

Table 2(18)

Tape Sample by Demographic Characteristics

Grade 11/Age 17

	TAPE 1	TAPE 2	TAPE 3	TAPE 4	TOTAL
TOTAL	1539	1540	1596	1534	6209
SEX:					
MALE	774	791	796	745	3106
FEMALE	765	749	800	789	3103
RACE:					
WHITE	1065	1079	1158	1130	4432
BLACK	263	242	258	193	956
HISPANIC	148	163	121	172	604
OTHER	63	56	59	39	217
REGION:					
NORTHEAST	300	325	336	327	1288
SOUTHEAST	403	407	405	379	1594
CENTRAL	453	391	443	459	1746
WEST	383	417	412	369	1581
PARENTS ED:					
LESS THAN HIGH SCHOOL	194	194	203	161	752
HIGH SCHOOL	537	558	564	543	2202
GREATER THAN HIGH SCHOOL	749	696	779	775	2999
UNKNOWN	59	92	50	55	256
STOC:					
RURAL	118	58	85	106	367
DISADVANTAGED URBAN	198	181	159	179	717
ADVANTAGED URBAN	249	190	123	221	783
BIG CITY	108	126	233	141	608
FRINGE	208	132	177	170	687
MEDIUM	240	249	253	240	982
SMALL	418	604	566	477	2065

Chapter 3

DEVELOPMENT OF THE YEAR 15 NAEP READING AND WRITING ASSESSMENTS

Ina V. S. Mullis

Educational Testing Service

In developing each subsequent assessment, NAEP has had the twofold responsibility of 1) measuring trends in achievement, and 2) using improved methods to measure current educational objectives. Because fulfilling the first part of this assignment is anchored in repeating past practices and the second part requires innovative new measures, accomplishing NAEP's dual goals requires ingenuity. Many conflicts arise naturally in developing unified assessments when consultants suggest evaluation approaches that are simply beyond the scope of NAEP's resources and capabilities. Thus, the development process for each assessment must be undertaken carefully--the process is akin to rebuilding a boat while keeping it afloat throughout the rebuilding. Yet, these kinds of dilemmas are familiar to NAEP, and new procedures reflecting the lessons of experience and future concerns were systematically introduced into the Year 15 reading and writing assessments.

Responsibility for developing the materials for the Year 15 assessment occurred primarily during NAEP's previous grant period when the project was administered by the Education Commission of the States. The decision to assess writing and reading in 1983-84 was made by the NAEP Assessment Policy Committee; this is one of their duties specified in the NAEP legislation. It should be further noted that the NAEP legislation also mandates assessment of reading and writing every five years; given a biennial assessment schedule, both subject areas had to be assessed in 1983-84 to comply.

Prior to Year 15 (1983-84), NAEP had most recently assessed writing in Year 10 (1978-79); however, that assessment was one of three conducted in that year. This meant that the Year 10 assessment had been relatively limited in scope; therefore, NAEP planned a much larger assessment in Year 15. As a result, staff knew the development task would be substantial and work began in March of 1981. In contrast, NAEP had conducted a very extensive combined assessment of reading and literature in Year 11 (1979-80). Thus, it was felt that fewer new materials would need to be developed for the Year 15 reading assessment. The development of that assessment began in the fall of 1982.

The following represent some of the major issues addressed by the previous NAEP staff in developing the new materials for the Year 15 reading and writing assessments:

- Although the decision to assess writing in Year 15 was made prior to the decision to also assess reading, once reading assessment was underway the development of the two areas was coordinated as much as possible. The Year 15 assessment was almost conceptualized as one single area with three components--reading, writing, and writing about reading.

- NAEP developed a large number of background questions, particularly in the area of writing. The panels expressed a strong desire to develop questions that would lead to a greater understanding of writing instruction, student writing practices, and student perception of the value of being able to write well.

- NAEP had routinely developed school questionnaires, but for the first time teacher questionnaires were developed to be administered to English teachers. The questions reflected areas emerging from the effective schools research as conducive to improved performance as well as those of special concern to reading and writing educators.

- NAEP had previously reported writing achievement results based on relatively few writing tasks at each age level. NAEP's technical advisory committees had expressed reservations about the generalizability of the results and urged staff to assess a greater number and variety of writing tasks at each age level.

- The high priority given to the issue of the writing process by NAEP panels was clear. A great deal of effort was expended trying to develop assessment methods to allow students to engage in the writing process without dictating that they use particular strategies. Given the parameters of NAEP procedures and capabilities, however, the final decision was made to assess students' use of the writing process in a limited fashion.

- As a result of the Year 11 combined reading and literature assessment, many of the reading comprehension passages were literary in nature. Given the importance of reading across the curriculum and in out-of-school situations, NAEP concentrated on a new design based on assessing reading using social studies, science, and literary passages, as well as functional materials.

- Based on previous assessments, NAEP was concerned that the reading assessment pool had very few materials that challenged 17-year-olds. Thus, an effort was made to develop more sophisticated measures of reading comprehension for use at that level.

The following two sections, the first describing the process used to develop the writing assessment and the second presenting details about the development of the reading assessment, provide further information about the procedures and the consultants used to develop and select the assessment items, the issues raised, and the final decisions regarding the Year 15 assessment tasks, background questions, and questionnaires.

3.1 Developing the Year 15 Writing Assessment

Prior to the Year 15 assessment, NAEP had assessed writing three times--in Year 1 (1969-70), Year 5 (1973-74), and Year 10 (1978-79). However, NAEP had published only two previous sets of writing objectives, one in 1969 and the other in 1972, with a brief supplement added for the third assessment in Year 10. For the fourth national writing assessment, the many new advances in writing education that had taken place by the early 1980s dictated a total recasting of the 1972 objectives rather than further modification. Therefore, in March of 1981, the Writing Advisory Committee met for the first time to accomplish several tasks crucial to the development of NAEP's Year 15 writing assessment. The first was to begin development of a draft of new writing objectives for the Year 15 assessment; the second was to develop a statement of recommendations for the design of the assessment; and the third to plan the development of the writing assessment items.

3.1.1 The Year 15 Writing Objectives

The booklet Writing Objectives, 1983-84 Assessment (1982) contains the names of the Year 15 Writing Advisory Committee and the numerous consultants who participated in developing those objectives.

The objectives for the Year 15 assessment are based on the premise that individuals generally write for a purpose and an audience. Some writing is personal, intended for oneself or perhaps an intimate friend, whereas other writing is more public and is intended to communicate ideas and experiences to others. These objectives distinguish between two different major purposes by describing the first under Objective I--Students Use Writing as a Way of Thinking and Learning--and describing the second under Objective II--Students Use Writing to Accomplish a Variety of Purposes. Objective I discusses the ways in which students may undertake personal kinds of writing as a way of improving thinking skills and of learning both subject knowledge and knowledge about themselves. Objective II deals with the types of writing students are more likely to do in school or social settings. Objective II presents three primary purposes for public writing: informative, persuasive, and literary. There are, of course, other ways to describe these purposes of writing, and earlier sets of NAEP objectives used somewhat different terminology.

One major shift in the focus of writing education has been from an emphasis on writing products to an emphasis on the writing process. Objective III, Students Manage the Writing Process--reflects this change in focus. To discuss the process, it is necessary to present its components as if they are discrete operations; however, in reality they are interwoven parts of the entire process and not readily separable in practice. The recursive nature of the writing process and the interdependence of the generating, drafting, revising, and editing skills it requires cannot be overemphasized. Objective IV--Students Control the Forms of Written Language--discusses control of such skills as organizing, elaborating and appropriately using the conventions of writing (usage and mechanics). Objective V--Students Appreciate the Value of Writing--underscores the importance of students' learning why writing is a valuable personal and social activity.

3.1.2 Writing Objectives Development

As can be seen from the following description of the objectives development process, a wide range of people interested in writing education participated in the creation of the writing objectives for the Year 15 assessment. Work was done primarily in conferences, conducted by NAEP staff and consisting of approximately five to eight external consultants who drafted, revised, or reviewed the evolving objectives document. To maintain continuity, these conferences usually involved one or more members of the advisory committee. However, the purpose underlying the series of conferences described below was to adhere to NAEP's consensus development process and involve people with as many different viewpoints as possible in the development of the objectives. Subject-area specialists, parents, classroom teachers, school superintendents, curriculum specialists, state writing assessment personnel, and school administrators were all involved. All of these contributors and reviewers were chosen to reflect the perspectives of people in various sizes and types of community, from many geographic regions, and from various racial/ethnic groups.

In March of 1981, the Writing Advisory Committee began its deliberations on the new objectives by discussing the National Council of Teachers of English (NCTE) statement on Standards for Basic Skills Writing Programs. This document and the input of the committee members resulted in an outline for the Year 15 objectives as well as drafts of supporting text. The committee reconvened in July of 1981 to review and revise the drafts written by participants in the earlier meeting and by committee members in the interim. Concerns voiced about the original draft centered around the fact that the writing process appeared as discrete categories when in actuality these processes are varied and nonlinear.

Based on the work of the July meeting, and concerns raised, staff worked to prepare a draft document. This document was reviewed by additional writing specialists at a meeting in early October, revised by staff and consultants based on that review and shared with state writing assessment personnel later that month. Finally, in late October a consultant group, including both advisory committee members and additional writing educators, met to complete the draft and try to respond to all concerns raised by previous reviews.

In November of 1981, the Lay Review Committee met to review the Year 15 Writing Assessment Objectives draft. Recommendations made by the lay reviewers included expanding the section that dealt with the value of writing and adding a section that would offer practical suggestions for the application of the objectives.

The Writing Advisory Committee met again in December of 1981 to continue discussion and revision of the objectives draft. Later that month the revised draft was reviewed by two groups of external reviewers--the first was a group of writing researchers and the second, a group of elementary, middle, and high school teachers. Many suggestions for the section about putting the objectives into practice were obtained at the latter conference.

In February of 1982, another group of consultants was convened comprising both advisory committee members and additional external writing consultants. This group discussed recommendations from the previous two conferences and developed a revised objectives draft. Further, the section on "putting the objectives into practice" was developed and incorporated into the draft. Teachers and curriculum specialists were invited to participate in a conference to review the latest draft. Staff revised the draft based on this review and, in March and April, sent the revised draft out for a mail review by approximately 30 consultants representative of a variety of backgrounds and perspectives. Staff addressed the concerns of the mail reviewers and shared the resultant draft with a group of curriculum and instructional superintendents and coordinators from across the country.

In May of 1982, a working group of consultants was convened to address the concerns of the curriculum review. The work of this committee was reviewed by a subsequent group of consultants in June. Then, in late June, the Writing Advisory Committee met to review the objectives draft and make final revisions. This draft was reviewed for bias during July and August by members of the National P.T.A. The final review of the objectives booklet by external consultants was conducted at an August conference. The objectives were then edited and published by the Education Commission of the States in late 1982.

3.1.3 Writing Exercise Development

Since the first objective, writing to learn, seemed nearly impossible to measure under the time, resource, and paper/pencil methodological constraints of current assessment procedures, the NAEP Writing Advisory Committee decided that writing tasks should focus on measuring performance in informative writing, persuasive writing, and literary writing. In view of the objectives and past assessment experience, staff and consultants decided to strengthen the practice of assessing several kinds of discourse on the grounds that students may be proficient in some kinds of writing but not in others. Although some of the same skills are involved in each kind of writing, NAEP results amply illustrate that there are challenges and strategies unique to each writing task.

In addition, some information would be collected about student ability to manage the writing process (Objective III). Controlling the forms of written language would be addressed by evaluating sample responses for organization, cohesion, syntax, usage, and mechanics; information about how students perceive writing and writing instruction would be collected using multiple-choice instruments.

In summary, the writing tasks developed for the Year 15 writing assessment were to measure student writing performance in the areas of informative writing, persuasive writing, and literary writing, with some tasks including opportunities for pre-writing and/or revision to gather information about student familiarity and success in engaging in the writing process. Information about the writing process and students' perception about the value of writing would be measured using multiple-choice scales.

Given this broad guidance from the Advisory Committee, NAEP consultants and staff began developing new exercises for the Year 15 writing assessment in March of 1981. A list of consultants who participated in the process is found at the end of the section on writing assessment development. Several factors contributed to the scope of the task. First, NAEP did not have many writing tasks available from the Year 10 assessment and in view of the concerns about generalizability of results expressed by the technical committees, NAEP was very eager to enlarge the coverage of variety of aspects of writing for the Year 15 assessment. In addition, the tasks in that assessment did not attend to students' managing the writing process, and NAEP was very interested in field testing many different formats for tasks that allowed students to engage in the writing process. The committee felt adamantly that the writing process is internalized and implemented differently by different people and should not be structured or regimented by the assessment situation. As it transpired, none of the formats was very successful in both allowing flexibility and "forcing students to provide specific evidence" that they had engaged in the writing process. Eventually the Writing Advisory Committee suggested collecting most of the information about students' use of the writing process through background questions and leaving traditional procedures for administering writing tasks in place. Finally, NAEP had very few background questions from the previous assessment, and the objectives emphasized the writing process as well as students' attitudes and values toward writing. In short, NAEP had planned an ambitious writing assessment for Year 15 and most of the materials to implement that assessment needed to be newly developed.

The first exercise development conference in March of 1981 focused on developing measures of students' attitudes toward writing and the value they placed on it. Development of writing tasks was initiated at an April conference. In May of 1981, three item writing conferences were held--two to develop writing tasks and one to develop background measures about writing instruction and the writing process. In June an exercise review conference was held and both cognitive and non-cognitive measures were reviewed and revised. Yet another exercise development conference was conducted later that month to address concerns raised by the review committee.

The Writing Advisory Committee met in July to review all the measures developed during the spring of 1981. Based on its advice and further direction, two more exercise development meetings were held, one in July and one in August, prior to preparing the clearance package and conducting the first field test.

In October and November of 1981, field tests of both the writing tasks and various types of background questions, 20 booklets total, were conducted at a variety of sites around the country. All of these items were reviewed by the Lay Review Committee. The results of the field tests were reviewed by the Writing Advisory Committee and by an exercise development review committee in December. At a January 1982 development conference, some items were revised on the basis of the December reviews and additional new items were developed. In February the entire pool of items was reviewed by a committee of teachers representing elementary, junior high and high school. The comments and suggestions from the teachers' review were addressed at an item review and development conference held in March.

In April and May, nine booklets of newly developed and revised items were field tested at five sites across the country. The results of these field tests were shared at a meeting with teachers in June, and a writing development conference was conducted later that month to revise materials based on the field tests and teachers' suggestions. The Writing Advisory Committee met in June and reviewed the existing pool of items. Major difficulties with the items focused on trying to increase the quality and length of the responses to the writing tasks and trying to find the vocabulary to ask students about their instruction and use of the writing process. These concerns precipitated two additional exercise review and development conferences held in the month of August.

Very substantial field tests were conducted in October of 1982: thirty booklets per age were tested in seventeen sites across the country. The results were reviewed, items revised, and a subsequent field test conducted in December at twelve sites across the country. The Writing Advisory Committee met in January of 1983 to review the entire pool of items and to make the preliminary selections of both the writing tasks and background questions for the assessment. Based on this selection, staff prepared the writing materials for inclusion in the clearance package of the non-cognitive items for the Year 15 assessment which ECS submitted to NIE on February 3, 1983.

Subsequent to the transfer of the NAEP project to Educational Testing Service, ETS staff and consultants used these materials as well as the cognitive items to select the writing items for inclusion in the Year 15 assessment. The selected writing tasks and guides as well as the background and attitude measures were further reviewed by subject matter specialists and editors, as well as for bias, according to standards established by Educational Testing Service. These materials became the second and final set of writing items submitted for clearance for the Year 15 assessment.

A complete description of the writing tasks eventually assembled, printed, and administered as part of the Year 15 writing assessment is found in Section 3.1.5.

3.1.4 Writing Exercise Development Issues

3.1.4.1 Definitions of Types of Writing Tasks

The decision to increase the number of informative, persuasive and literary writing tasks raised an issue of considerable consequence--domain definition. Generally, mathematical operations have been well defined and considerable effort has been devoted to describing the specifics of science, social studies, and reading. In contrast, relatively little had been done to classify subtasks within purposes for writing. This problem, of course, could not be tackled in its entirety. However, enough progress was made to create writing task development frameworks and provide item writers with specifications. An overview of these frameworks follows.

Informative Writing (Objective II A)

Briefly, writing to inform others can involve reporting and retelling events or experiences. It can also involve analyzing or examining concepts and relationships or developing new hypotheses or generalizations from existing records, reports, and explanations. Tasks developed to measure informative writing can range from simple note taking and recounting events to explaining concepts and supporting generalizations, with particular attention to a balance between lower-level, or reporting, tasks and those tasks requiring higher-level, or analytic, skills.

Writing tasks requiring informational writing were designed to cover a range of difficulty levels, a range of audiences, a range of stimulus materials (including personal experience and given materials) and a variety of writing situations. The two major classifications were reporting and analysis, with subclassifications of each. For example, in the area of reporting: a note about where the student went after school; a letter of complaint; instructions on how to feed a pet; and a job application represented various functional writing tasks. Writing reports from diagrams or notes represented the kinds of writing tasks that may be required in school or business situations. In addition to these two contexts for writing, students were asked in some tasks to write on the basis of given information and in others to write based on personal experience.

The analysis tasks also were designed to give respondents the opportunity to use both personal experience and given material as the basis for presenting and supporting their ideas. These tasks were designed to measure higher-order skills by requiring respondents to advance from reporting facts to providing explanations. Again, there was an effort to represent both school and non-school contexts.

Persuasive Writing (Objective II B)

Persuasive writing may entail responding to requests for advice by giving an opinion and supporting reasons. However, it usually involves initiating an attempt to convince readers by setting forth one's own point of view with evidence to back it up. Argument, with refutation, becomes part of persuasion when the writer knows there is opposition to what he or she is advocating. Thus, persuasive writers must be concerned with the positions, beliefs, or attitudes of particular readers and with the possibility of winning their support or changing their beliefs or attitudes.

Tasks designed to measure persuasive writing capabilities included items carefully constructed to range from advice-giving to refutation. More specifically, "convince" items required students to give an opinion and the supporting evidence that would sway a particular audience; and "refute" items required students to take a position contrary to that of their audience and to give evidence that would advance their position and refute the expressed concerns of their audience.

Literary Writing (Objective II C)

Literary writing provides a special way of sharing experiences and understanding the world. There are a wide variety of forms that literary writing can take, such as stories, poems, plays, or lyrics. However, given the context of the assessment, the panels decided to focus the development of literary writing tasks in the area of storytelling. Tasks requiring both imaginative narratives and personal experience narratives were developed to offer students opportunities to write from a basis of imaginative ideas and a basis of their own experience. Also, several tasks were developed which asked students to attempt modest poems. Given the resource constraints of the actual assessment, the final selection reflected only imaginative narratives.

3.1.4.2 Prior Knowledge Bias

The controversy here was concerned with how best to reduce the effects of prior knowledge on performance without adversely affecting the level of that performance. The optimum strategy in designing "fair" writing tasks requires respondents to have equal levels of prior knowledge about each topic, where equal level can be defined as ranging from little or no knowledge to extensive knowledge. However, it is also agreed that more effective writing is produced when authors have extensive knowledge or at least some familiarity with the subject. Unfortunately, NAEP collects information from a national sample of students and universally appealing topics are extremely rare. Recently, even such traditional stimulus standbys as pets, vacations, and basic emotions have become suspect. Thus, the effort to reduce bias led ironically to writing topics that very few people were likely to know or care about; eventually, an insidious banality pervaded the entire item pool. This was, of course, troubling, since students could not possibly be inspired to do their best writing with such bland prompts. The solution was to compromise--some universal topics and some obscure topics, some based on personal experience and some on given material, some rural and some urban, some for girls and some for boys, etc. Thus, when achievement is summarized across the total set of topics, chances for better performance should be maximized, while bias is minimized.

3.1.4.3 Should Audience be Specified in Tasks?

Must a writing assignment specify an intended audience? Some writing task developers said yes, some said no. The "no audience" argument is that the respondents know their papers will be read by teacher-like graders. Therefore, specifying an audience other than a NAEP reader brings an artificiality to any task that will jeopardize performance. At the other extreme, and equally adamant, were those who insisted that it is impossible to do all on a writing task unless a specific audience is identified. Once again the consensus process yielded a compromise. The persuasive tasks delineate audiences to create a context for the persuasion, while the literary items rarely specify an audience on the grounds that the writer is frequently his or her own primary audience. Some informative tasks have audiences, whereas in others the audience is left to the imagination of the writer.

3.1.4.4 How to Assess the Writing Process

In dealing with this issue, NAEP was not so much faced with reconciling opposite points of view espoused by different writing experts as with reconciling paradox expressed by almost every advisor. The importance of focusing writing instruction on the writing process was clear. Advisors, item developers, and staff desperately wanted such measures. However, it was unanimously agreed that each person writes best when allowed to engage in the process as they have found it most effective. This dilemma was exacerbated by the knowledge that students have not done well in past assessments when given opportunities to revise, as they do not appear to know what to do. Therefore, on one hand, it was deemed necessary to give students some help by specifying the steps they should engage in to accomplish the process; on the other hand, forcing students through various steps without allowing for flexibility was considered detrimental. Unfortunately, given the parameters of NAEP procedures and capabilities, there was not a satisfactory solution to this measurement problem. Therefore, a very few items were developed that attempted to measure aspects of the writing process. Due to resource limitations, none were included in the assessment, although one successful item type required students to rewrite and improve a given piece of writing, rather than their own writing. Some of these included suggested directions for improvements. Field tests indicated that students may be more successful with these tasks than they are when asked to revise their own writing. Finally, numerous questions were developed that asked students about the prominence of the writing process in their instruction and whether they engage in various aspects of the writing process or utilize particular strategies when they write.

3.1.4.5 Summary of Writing Task Development Issues

Thus far this chapter has summarized NAEP discussions about some writing evaluation issues raised during the development of the fourth writing assessment. What follows outlines a somewhat expanded overview of the problems faced. The NAEP resolutions described were often not the preferred resolutions, but represented compromises based on the reality of assessment capabilities and resources.

(1) Should writing be assessed solely by collecting writing samples, or should some less costly multiple-choice or short-answer items be used?

NAEP resolution: Only writing samples should be used. They increase the utility of results, in that they appear more valid and each sample can be evaluated from a variety of perspectives.

(2) Should student performance be described by providing detailed information about a small number of tasks or by providing more general information about a wide variety of tasks?

NAEP resolution: Try to do both. Increase the number of writing tasks to provide better information about the range of tasks students can perform, but retain the capability to provide detailed information about some tasks.

(3) What kinds of writing tasks should be included in the assessment?

NAEP resolution: Informative tasks ranging from note-taking to analysis, persuasive tasks ranging for advice-giving to refutation, and literary tasks including a range of narratives. All tasks should try to be representative of naturally occurring writing situations and contexts both in and out of school.

(4) Should audience always be specified in writing tasks?

NAEP resolution: Only in persuasive writing tasks. Informative and literary tasks may or may not have audience specified. Further, any specified audience must appear natural, not artificial.

(5) How can NAEP address the prior knowledge issue?

NAEP resolution: Have a number of the tasks based on given information. For those tasks based on students' own experiences, avoid biased items while being sure to maintain a balanced pool.

(6) How can NAEP measure students' ability to engage in and manage the writing process?

NAEP resolution: Only in limited ways; perhaps in the next assessment.

(These, of course, are not the only issues raised during the course of developing the fourth national assessment of writing.)

3.1.5 NAEP's Year 15 Writing Assessment Exercises

Informational Writing--Reporting

From Personal Experience

Ages 9, 13/Grades 4, 8	Pets: Students were asked to write a note explaining to a friend how to care for a pet while they were away on vacation, including where to find the food, how often to feed the pet, and how much food to give the pet.
Age 17/Grade 11	Job Application: Students were asked to provide a brief description of a desirable summer job and to describe the experiences or qualifications they had for such a job.

From Given Information

Age 9/Grade 4	Plants: Students were asked to summarize a science experiment based on a series of pictures of different stages of a plant's growth.
Ages 9, 13, 17/Grades 4, 8, 11	Appleby House: Students were asked to write a newspaper article based on notes provided about an unusual haunted house.
Ages 9, 13/Grades 4, 8	XYZ Company: Students were asked to send away for a T-shirt in response to an advertisement.
Ages 9, 13, 17/Grades 4, 8, 11	Dali: Students were asked to describe a surrealistic painting by Salvador Dali.

* * *

Informational Writing--Analytic

From Personal Experience

Ages 9, 13, 17/Grades 4, 8, 11	Favorite Music: Students were asked to describe a favorite type of music and explain why they liked it.

From Given Information

Ages 9, 13, 17/Grades 4, 8, 11	Food on the Frontier: This task began with a passage about frontier life; students were then asked to compare modern-day food with frontier food.

* * *

Persuasive Writing--Convincing Others

Age 9/Grade 4	Spaceship: Students were asked to argue for permitting captives from outer space to return home rather than detaining them for scientific study.

Age 13/Grade 8	Dissecting Frogs: Students were asked to discuss and support their views on dissecting frogs in science class.
Age 17/Grade 11	Space Program: Students were asked to take a stand on whether funding for the space program should be cut, and why.
Ages 13, 17/Grades 8, 11	Split Session: Students were asked to write a letter requesting a morning or afternoon school session and explaining their preference.
Ages 9, 13, 17/Grades 4, 8, 11	Swimming Pool: Students were asked to write a letter to a swimming pool manager, convincing the person to hire them for a summer job at the pool.
Ages 9, 13, 17/Grades 4, 8, 11	School Rule: Students were asked to express a desire for changing a school rule and to discuss why.

Persuasive Writing--Refuting an Opposing Position

Age 9/Grade 4	Aunt May: This task asked students to write a letter convincing Aunt May they are old enough to travel alone even though Aunt May thinks otherwise.
Ages 9, 13/Grades 4, 8	Radio Station: Students were asked to give reasons why their class should be allowed to visit a local radio station despite the manager's concerns.
Ages 13, 17/Grades 8, 11	Recreation Opportunity: Students were asked to take a stand on whether a railroad track or a warehouse should be purchased and to argue on the basis of possible recreational opportunities.
Age 17/Grade 11	Uncle: Students were asked to write a letter to an uncle convincing him to lend his car so the student could visit a friend. Responses needed to explain the situation, convince the uncle that the student was a safe driver, and to do so without hurting the uncle's feelings.

Age 17/Grade 11	Bike Lane: Students were asked to take a stand on whether a bike lane should be installed and to refute specific opposing views.

* * *

Imaginative Writing

Ages 9, 13, 17/Grades 4, 8, 11	Hole in the Box: Students were given a picture of a box with a hole in it and an eye peeking out; they were asked to imagine themselves in the picture and then to describe the scene and how they felt about what was going on around them.
Ages 9, 13, 17/Grades 4, 8, 11	Ghost Story: Students were asked to write a good, scary ghost story.
Ages 9, 13, 17/Grades 4, 8, 11	Flashlight: Students were asked to write a story about adventures with a flashlight with special powers.

3.1.6 Background Questions

In NAEP's attempt to trace the effects of instructional practices on student performance, the Year 15 writing assessment included more non-cognitive student background questions than ever before. These focused on the students' attitudes toward writing, the strategies they used to complete their writing assignments, the kinds of writing they did in school, and the kinds of instruction and help they reported that they had received from their teachers.

Because this is an era in which schools across the country have increased the priority they place on writing instruction, both in the kinds and amounts of writing students are asked to do in school and in the kind as well as amount of help they receive from their teachers, it seemed particularly timely to describe students' perceptions of their instructional environments and to relate these to writing proficiency. Over 100 background questions specific to writing were included at each age level. Details of the non-cognitive assessment are included in Chapter 6.

3.1.7 Evaluation of Student Responses to Writing Tasks

Throughout the winter of 1982-83, conferences were held with consultants to develop and refine primary trait scoring guides and document them with illustrative sample papers from the field tests. The primary trait scoring method reflects students' success in accomplishing the specific informative, persuasive, or imaginative writing task. Primary trait results for accomplishing the task are based on levels of success. Responses are either rated as **unsatisfactory**, **minimal**, **adequate**, or **elaborated**, or they are **not rated**. Although criteria for the categories are specified in terms of each writing task, a general explanation of these levels follows.

Levels of Task Accomplishment

Not rateable. A small percentage of the responses were blank, indecipherable, totally off task, or contained a statement to the effect that the student did not know how to do the task; these responses were considered **not rateable.**

Unsatisfactory. Students writing papers judged as **unsatisfactory** provided very abbreviated, circular, or disjointed responses that did not represent even a basic beginning toward addressing the writing task.

Minimal. Students writing at the **minimal** level recognized some or all of the elements needed to complete the task, but did not manage the elements well enough to assure the purpose of the task would be achieved.

Adequate. **Adequate** responses included the information and ideas critical to accomplishing the underlying task and were considered likely to be effective in achieving the desired purpose.

Elaborated. **Elaborated** responses went beyond the essential, reflecting a higher level of coherence and providing more detail to support the points made.

In addition to being evaluated in terms of task accomplishment, student responses collected to measure trends in performance across assessments were rated holistically to provide an overall estimate of the relative fluency of the writing. Readers did not make separate judgments about a paper's organization, content, grammar, usage, spelling, and punctuation, but judged the overall effect of the paper. In contrast to the evaluations for task accomplishment, where responses to the same task written by more than one age group were evaluated against the same specific criteria, fluency was evaluated by rating papers on general impression relative to other papers from the same age group. (For example, a response to a given task written by a 9-year-old was ranked in comparison to the responses written by other 9-year-olds in the Year 15 as well as previous assessments.) Each response was given a rating from the highest to the lowest according to six levels of fluency, with six being highest.

Overall quality measures are complemented with information about syntax and mechanics. A syntactic analysis involves breaking up each paper into "T-units" (an independent clause and all of its modifying words, phrases, and clauses) and examining the ways in which writers embed information in T-units and join T-units together. A mechanics analysis involves classifying the kinds of errors writers make in sentence use, punctuation, spelling and so forth.

3.1.8 Writing Exercise Development Consultants*

Arthur Applebee
Stanford University
Stanford, CA

David Bartholomae
University of Pittsburgh
Pittsburgh, PA

Elsa Bartlett
New York University Medical Center
New York, NY

Bill Burns
Boulder High School
Boulder, CO

Courtney Cazden
Harvard University
Cambridge, MA

Jane Christiansen
National Council of Teachers of English
Urbana, IL

Charles Cooper
University of California
San Diego, CA

John Daly
University of Texas
Austin, TX

Vivian Davis
Tri-Ethnic Committee
Dallas, TX

Paul Diehl
University of Iowa
Iowa City, IA

Marjorie Farmer
Philadelphia Public Schools
Philadelphia, PA

Ed Folsom
University of Iowa
Iowa City, IA

Donald Graves
University of New Hampshire
Durham, NH

Robert Gundlach
Northwestern University
Evanston, IL

Kris Gutierrez
University of Colorado
Boulder, CO

Diane Hernandez
Lafayette Elementary School
Lafayette, CO

---

* See Writing Objectives, 1983-84 Assessment (1982) for a list of consultants who participated in developing writing objectives.

Ann Humes
Southwest Regional Labs
Los Alamitos, CA

Don Jones
Jefferson County Schools
Lakewood, CO

Kenneth Kantor
University of Georgia
Athens, GA

Carl Klaus
University of Iowa
Iowa City, IA

Judith Langer
Bay Area Writing Project
University of CA
Berkeley, CA

Richard Lloyd-Jones
University of Iowa
Iowa City, IA

Carol Mathews
Boulder High School
Boulder, CO

George McCulley
Michigan Technological University
Houghton, MI

Mary Meier
Eugene School District
Eugene, OR

John Mellon
University of Illinois
Chicago Circle
Chicago, IL

Patti Mendes
University of Colorado
Boulder, CO

Jeff Oliver
Lincoln Elementary School
Boulder, CO

Jesse Perry
San Diego Public Schools
San Diego, CA

Anthony Petrosky
University of Pittsburgh
Pittsburgh, PA

Edys Quellmalz
Center for the Study of Evaluation
University of California
Los Angeles, CA

Sandra Seale
Cherry Creek High School
Englewood, CO

Mary Ann Shea
University of Colorado
Boulder, CO

Yvonne Siu-Runyan
Boulder Valley Education Center
Boulder, CO

Susan Sowers
Harvard University
Cambridge, MA

Cary Stitt
Jefferson County Schools
Lakewood, CO

Lynn Troyka
City University of New York
College Rayside, NY

Tomas Vallejos
University of Minnesota
Minneapolis, MN

Faith Waters
Bucks County School District
Doylestown, PA

Darnell Williams
Bishop College
Dallas, TX

John Wood
Juchem Elementary School
Broomfield, CO

3.2 Developing the Year 15 Reading Assessment

Prior to the Year 15 assessment, NAEP had completed three assessments of reading and one of reading and literature combined. The first assessments of reading and of literature were in Year 2 (1970-71). Reading was re-assessed in Year 6 (1974-75) and Year 11 (1979-80) using a subset of the reading items for the first assessment. Literature was re-assessed in Year 11 using a few items from the first literature assessment. Also during Year 11 reading and literature were assessed together using a new, combined set of items. This document summarizes the design and development of the Year 15 assessment of reading including revision of the reading objectives, exercise development, field testing and exercise reviews.

3.2.1 NAEP's Year 15 Reading Objectives

The Year 15 reading assessment was developed to address four major objectives (see the booklet, Reading Objectives, 1983-84 Assessment [1984]). The first objective, **Comprehends What is Read**, is central since every other objective is an outgrowth of that one. It includes comprehension of various types of written materials read for a variety of particular purposes. The second objective, **Extends Comprehension**, includes analyzing, interpreting and evaluating what has been read.

Good readers develop a variety of strategies to help them comprehend what they read. The third objective, **Manages the Reading Experience**, addresses how a reader might adopt various strategies depending upon the characteristics of particular passages, the reader's knowledge and experience with similar materials, and the reader's purpose for reading. The fourth objective is **Values Reading**.

3.2.2 Reading Objectives Development

NAEP had expended considerable effort developing the Reading and Literature Objectives, 1979-80 Assessment which was published in 1980. The eighteen-member Reading/Literature Advisory Committee guided the development of those objectives and approximately 130 consultants participated in the development process. Given the extent of this effort and that only two years had passed, NAEP explored the possibility that the Year 11 objectives might be viable for the Year 15 assessment. To review the appropriateness of the Year 11 reading/literature objectives for the Year 15 assessment, the Year 11 objectives were mailed to a group of reading and literature experts who were asked to comment on any additions or changes they felt should be incorporated into the objectives.

Second, the objectives were discussed by a smaller group of consultants at a meeting held in Denver on December 9-11, 1982. The group reviewed the comments from the mail review (without knowing specific authors), discussed their own comments, and then reached consensus regarding their recommendations. The group generally approved of the content of the objectives, but felt the objectives should be rewritten in a clearer style; more specifically, the group made six recommendations:

(1) The concept of proposition needed clarification in the comprehending objective. A complete rewrite was drafted at this meeting and the word "proposition" was eliminated. The concept of amount of text, characteristics of text, prior knowledge required and the overall interactive nature of the comprehending process was addressed. The decoding aspect of reading comprehension was mentioned only as a prerequisite and was not identified as an area to assess.

(2) The responding objective needed to be expanded to include non-literary texts. A complete rewrite of this objective also was drafted at this meeting. An effort was made to be sure the connection between the comprehending and responding objectives (all part of the same process) was clear at the onset. The cognitive and emotional aspects of the comprehending and responding objectives were presented as being interrelated.

(3) The valuing objective needed expansion in several ways. The descriptions of valuing reading as a source of pleasure and the obtaining self-understanding needed embellishment. Mention of the value of reading in gaining practical knowledge was seen as necessary. A section discussing the fact that valuing reading is not a goal for all cultural groups and that different cultural groups gain value from reading in different ways also required mention. The section on the cultural role of reading needed to be more explicit about freedom to publish and freedom of access to published material.

(4) The study skills objective needed to be edited.

(5) A section that deals with issues related to text needed to be added. This section should include discussion of how text structure and prior knowledge affect reading, and issues of text selection. Also, the variety of text types should be discussed with a mention of the importance of practical reading. The section would not imply that skills of metacognition should be assessed.

(6) The consultants urged that a major section dealing with instructional implications be added.

These suggestions were incorporated into a new draft of the reading/literature objectives.

This draft was reviewed by additional external consultants and revised by NAEP staff and consultants. The resultant copy was subjected to a lay review by persons involved and interested in education. The comments and suggestions of the reviewers were addressed by staff and this final draft of the Reading Objectives was subsequently edited and published by Educational Testing Service in 1984. Reading Objectives, 1983-84 Assessment (1984) lists the participants in the objectives development process.

3.2.3 Reading Exercise Development

Development of the Year 15 reading assessment began in the fall of 1982. The design included a new approach: assessing reading within the content areas of literature, science, social studies and a few "out of school," or media and functional reading materials likely to be encountered by students in their day-to-day experiences. The design for the Year 15 assessment also called for the separate assessment of reading and writing skills as well as joint assessment of these skills. This design was an expansion of the model used in the Year 11 reading assessment where students read a passage, answered multiple-choice comprehension questions and then wrote about the passage. The new approach seemed particularly appropriate for measuring Objective II--analyzing, interpreting, and evaluating what has been read. It also reflected the new emphasis on integrated language arts instruction and assessment, and application of language skills in the various school content areas and everyday life tasks.

3.2.3.1 Passage Selection

Selecting reading passages to use as stimuli for the reading comprehension questions was the first step of the development process. Educators and reading professionals were asked to select passages according to guidelines developed from NAEP's experience with past assessments. The guidelines, outlined below, dealt with the use of the stimulus materials, their length, formats, possible sources and general review criteria. The consultants were asked to concentrate on selecting passages in their areas of expertise--literature, science, or social studies, and elementary or secondary school levels. (The consultants who participated in item development are listed in Section 3.2.5)

Passage Selection Guidelines

> Use of Stimulus Materials. It is essential to keep in mind that the materials you are selecting will be used as stimuli for assessment items. It is important that they contain information, problems, characters, situations that will supply information for developing reading comprehension and writing items.
>
> Length of Stimulus Materials. The materials that you pick need to be relatively short. We do not want to have the reading process to take up too much assessment time. A guideline is to keep all materials within the limit of 30 to 2000 words with very

few that are long. Ideally, NAEP needs short but substantive stimuli.

Formats for Stimulus Materials. In the vast majority of the cases NAEP needs normal linear text material for stimuli, i.e. material presented in full sentences and paragraphs. However, in a few cases some non-linear written material may be appropriate, e.g. charts, graphs, tables, advertisements, application forms, and so forth. These non-linear materials will probably be most applicable for the single theme modules or for the out-of-school materials that are explained later.

Sources for Stimulus Materials. Select stimuli from existing published material, but not from widely distributed curriculum series. Good sources of materials are supplementary curricular materials and modules, material from resource books, educational magazines and newspapers, and so forth. You may use excerpts from materials. However, the material you select must stand alone as a complete piece.

Criteria for Reviewing Stimulus Material. Some overall criteria for reviewing stimulus material are provided in Table 1 (see below). These are the same criteria that we will be using as we review the pool of materials and that we will give to outside reviewers. Keep these criteria in mind throughout your search for materials.

Type of Stimulus Materials. The **literature** materials may include all types of fiction material, e.g. stories, poems, plays, etc. as well as nonfiction material that has a literary quality, e.g. a vivid character sketch that describes a real person. **Social studies** passages should be typical of materials that students read as a part of their social studies curriculum. However, as indicated earlier, do not include material from widely used curricular material. In selecting materials, assume that students have some basic content knowledge; do not limit yourself to the very simplest instructional materials. Try to select some stimulus materials that reflect the special interest of women, blacks, Hispanics and other minority groups. The guidelines for the **science** materials are similar to those for social studies. Finally, include some materials which are more typical of **out-of-school** reading. Two specific types of material are sought: functional and media. Functional materials may include such things as instructions, labels, forms and so forth, that students have to deal with in their everyday functioning. Media materials are those that are typical of newspapers, magazines, posters, radio and television. They may include news stories, editorials, advertisements, commercials, and so forth.

TABLE 1
CRITERIA FOR REVIEWING STIMULUS MATERIAL

(1) Naturally occurring

- -- Stands on own, not segmented
- -- Reflects materials that students commonly read in or out of school

(2) Interesting for target group(s)

- -- Reflects the topics and genre which target age group(s) enjoys reading

(3) Relevant to experience of target age groups(s)

- -- Reflects settings and activities which students from all parts of the country, racial/ethnic backgrounds and economic backgrounds find reasonably familiar

(4) Appropriate difficulty level for target age group(s)

- -- Appropriate vocabulary level and readability following general guidelines for difficulty

(5) Meets at least minimum standards of writing

(6) Not offensive and/or stereotypic

- -- Shows wide variety (traditional and non-traditional) of roles, personalities, etc.

(7) Enduring

- -- Relevant for future assessments (1990)

(8) Not widely read in regular school programs

3.2.3.2 Passage Review

A conference to review and select from the pool of passages received was held October 7-9, 1982. The group of reviewers was composed of reading and content area experts representing different geographical, cultural and ethnic backgrounds. The meeting began with a general orientation and review of the passage selection guidelines. After initial orientation, the group broke up into three small groups--literature, science and social studies groups. Each group read through their set of passages, coming up with a consensus rating of excellent, good, fair or unacceptable for each passage. A final pool of passages was created from those rated as good and excellent. These passages were then reviewed as a set, determining which particular areas were weak in quality or representation. All three groups felt that certain areas needed more coverage, and that higher quality materials could be found. Both the literature and social studies groups felt that they could find better passages with female and minority characters and themes. The social studies group felt that global issues and urban themes were under-represented in their pool of passages. The science group felt that better passages for 17-year-olds were needed. The consultants were asked to select new passages for these areas.

3.2.3.3 Exercise Writing

In October of 1982, all passages selected by the reviewers were sent to educators and reading and measurement specialists. These consultants were provided with instructions, guidelines and examples for writing exercises to accompany the passages; an overview of these guidelines is presented below. They also attended one of two meetings (one in NAEP offices in Denver, the other at the University of Illinois at Champaign) for training in exercise writing. The exercises received from these consultants were then reviewed, edited and revised by NAEP staff members and an external consultant. An item documentation system was developed to track passages, items and review ratings.

Exercise Development Guidelines

> In the item specifications, we have provided you with many suggested item stems, for example, "What is the main idea of the story?" and many examples of items from the previous assessment. These guidelines are meant to be suggestive, not mandatory. The nature of the passage will often dictate different ways of asking a particular question, or different questions altogether. Let the text and the obvious important areas of meaning guide the types of questions you ask and the way you ask them.
>
> You will notice in the item specifications and examples that we sometimes have different ways of asking the same question for different ages. For example, we ask 9-year-olds "How does the writer make the story sound?" and we ask 13- and 17-year-olds "What is the tone of the story?" These are essentially the same

question, but we have attempted to make the question easier for 9-year-olds. Please make these same kinds of modifications when you are writing questions for 9-year-olds.

Item Difficulty. It is very difficult to draw a line between making item distractors plausible and discriminating between good and poor comprehension and making item distractors unnecessarily tricky and misleading. There are two minimum criteria to follow and after that it is a matter of judgment. First, both item stems and distractors should use vocabulary and syntax that are easy to understand. The student should be tested on his or her comprehension of the passage only, not his or her comprehension of the question. Second, each question should have a single correct answer, one that can be clearly defended. Beyond this, we would prefer to err on the side of making items difficult rather than easy, especially for ages 13 and 17. Our current pool of assessment items is too easy at the older ages. This problem is partly due to passages that are easy for older students and partly due to items with options that are very obviously correct or incorrect.

Background Knowledge. Another difficult judgment has to be made regarding the amount of background knowledge that is required to answer a question. Background knowledge plays an important part in the comprehension process. For many comprehension tasks the student must bring some of his or her own knowledge or experience to the passage in order to answer a question. While writing comprehension items, you need to keep in mind two possible extremes: 1) items that are based on background knowledge that all students have; 2) items that are based on background knowledge that very few students have or that only students from a particular group have.

With respect to the first point, you should not write an item that many students could answer without reading the passage. For example, you should not ask a question, "Who was George Washington?" even though the passage was about George Washington.

With respect to the second point, you should not write items that are based on very specialized knowledge or on knowledge that only special group of students might have. For example, you should not write items that require detailed knowledge about the Middle Ages or detailed knowledge about life in rural areas.

Sometimes you may be faced with a situation where the passage provides information that might conflict with information that the student may already have about a topic. In order to reduce confusion you may preface the question with "According to the article...." However, we suggest you use this method sparingly. If the item you write produces a lot of dissonance between the information in the text and general knowledge, it probably is not a good item.

Irrespective of the problems presented by background knowledge, you should not avoid writing questions that require some background knowledge altogether. Background knowledge is a very important requirement for higher level reading comprehension and cannot be avoided in the writing of reading comprehension items.

Number of Types of Questions. For each passage, you should write questions based on the important meanings that can be derived from the text. The number of questions will vary depending upon the length and the content of the passage. Some passages are very short and only present a few ideas. Some are fairly long and present a very large number of ideas. As a guide, we would like you to write 3 or 4 items for short passages (less than 200 words) and 6 to 8 questions for longer passages (200 or more words). We want more items than we will ultimately use in the assessment because we know that many will be deleted during the field-testing and review processes. However, don't struggle to meet a quota. Write as many items as you think it takes to cover the major meanings of the passage. For longer passages you will only be able to write items that sample the major meanings of the passage.

Please document each question, by indicating the specific reading comprehension task it measures and by providing a rationale for why that task is an important one.

Passage Modifications. The passages that we provide you with all come from actual published documents. We will gain permission from their publishers to use them in the assessment.

Although NAEP wishes to maintain the natural quality of the passages, slight modifications are possible. You may receive a long article and wish to shorten it. You may find that the text needs a word changed or an introductory sentence added. In some cases, you may need to add an advanced organizer to a passage in order to give the student some background information. However, we suggest you use advanced organizers sparingly.

3.2.3.4 Bias Review

In early December, 1982, the reading and writing passages and exercises were sent to consultants representing various constituent groups (e.g., minorities, women's groups, large urban school systems, academicians). These consultants reviewed, rated and made recommendations for improvements in the passages and exercises to be used in the assessment. The guidelines sent to these reviewers follow.

General Guidelines for Judging Bias

Items should reflect settings and activities that are reasonably familiar to students from all regions of the country, regardless of economic or racial/ethnic background.

Items should not be offensive or stereotypic to any segment of the population. Typical kinds of stereotypic descriptions to be eliminated are those involving: sex, race, culture, ethnicity, older persons or handicapped persons.

Other stereotypic descriptions to be avoided might involve the following: social roles, psychological traits, physical appearance, occupation, life style or language.

Other possible reasons for classifying passages or items as unacceptable are:

- language, descriptions or situations presented which might be offensive to any segment of the population;
- passages or items that may not be within the realm of experience of students from particular geographic areas or socio-economic situations;
- words that may not have a common meaning for everyone; and
- exceptionally difficult or complex vocabulary or sentence structure.

Please use a separate rating line for each separate passage. If a passage has several questions accompanying it, please rate all items for that passage on the same line, indicating by part letter any that are rated differently. If all items associated with a passage are acceptable, no separate listings by part are required.

3.2.3.5 December Field Testing

New passages and exercises were packaged to be field tested at a variety of sites across the country in early December.

The passages and exercises were packaged into twenty booklets, seven for 9-year-olds, seven for 13-year-olds, and six for 17-year-olds. For all booklets, the exercises were ordered to vary in length, type of passage, type of exercise and difficulty. Booklets began with easy items, longer and more difficult items were placed in the middle of booklets along with open-ended writing exercises, and shorter, easier items were placed at the end.

3.2.3.6 Exercise Review

A meeting was held January 27-28, 1983 to review results obtained from the December field tests. A group of reading, measurement and curriculum specialist was given a set of the items to review before the meeting, as well as guidelines for review. At the meeting, the group was provided with field test results and comments from bias reviewers.

Consultants were first given an orientation to the reading assessment and the criteria for exercise review and selection. They then worked in small groups, again concentrating on the areas of literature, science and social studies. They rated each reading passage as very good, good, fair or unacceptable. Consultants also rated each question as acceptable, acceptable with modification (suggesting specific modifications) or unacceptable.

In large group debriefing, the consultants discussed their feelings about the pool of items selected. The literature group felt the literature passages were greatly improved over those used in the Year 11 assessment, needing only an addition of one Hispanic-oriented passage for 9-year-olds. The science and social studies groups felt that the passages in these two subject areas did not represent typical textbook material. Also, many of the science and social studies materials lacked overall coherence--lacking major premises, supporting examples or summaries. Once again these consultants were asked to find new passages to remedy these problems.

3.2.3.7 New Selection and April Field Testing

From February through March, a small number of new science and social studies passages were identified and items developed. Also, recommendations from the passage reviewers were incorporated into the earlier items. New and revised items were then subjected to a review by NAEP staff members and an external consultant. Reading items were packaged to be field tested April 24-29, 1983. Some exercises that had survived previous reviews were dropped or changed based on recommendations by Educational Testing Service. A total of eighteen test booklets were produced: five at age 9, seven at age 13 and six at age 17. A small number of items were overlapped at two ages, and one item was used for all three ages. The items were packaged in booklets with accompanying answer sheets. The answer sheets provided space for students to record the time after they completed reading a passage and answering questions for that passage.

In addition, a number of background questions for students, teachers, school administrators and principals were developed and field tested. These questions were directed at variables that have impact on student achievement. More specifically, NAEP field tested a school characteristics and policy questionnaire concerning use of principals' time, incorporation of results of school effectiveness research into the school, school climate and school improvement. The teachers' questionnaire asked about resources, preparation and training, instructional objectives, instructional practices, materials, evaluation techniques and school climate. Students' questions asked about activities and preferences, their study and library activities, and their classroom experiences.

Staff and consultants reviewed the field test results and reviewed the entire set of newly developed and previously assessed reading items. A selection was made for the Year 15 assessment and the background materials, including the school and teacher questionnaires, were prepared for inclusion in the clearance package. ECS submitted the clearance package containing student background questions for reading and writing, the English teachers' questionnaires, and the school characteristics and policy questionnaires on February 3, 1983.

3.2.4 The Year 15 Reading Assessment Exercises

The final review and selection of the items for the Year 15 reading assessment was conducted by staff and consultants at Educational Testing Service. All the items were reviewed by subject matter specialists, measurement experts, and editors as well as for bias according to the ETS Standards for Quality and Fairness. A second clearance package containing both cognitive and non-cognitive items was submitted for OMB clearance. These materials became the basis for assembling the reading blocks for the Year 15 assessment.

The Year 15 reading assessment materials included a variety of tasks and a variety of stimulus materials and, therefore, represented a range of topics and difficulty. Students were asked to respond to multiple-choice questions, to answer brief open-ended questions, and to write about their reactions to what they read. Short and long passages, graphically presented materials, poems, "real-world" materials, and reference materials were all included in the assessments. The majority of the materials were drawn from those developed for previous assessments. To measure trends across time, one group of exercises had been used in three previous assessments and a second group of exercises had been administered once before, in Year 11. To contribute to a more complete picture of current levels of reading performance, new exercises developed specifically for the Year 15 assessment were also included. The new exercises reflect an increased interest in students' abilities to read "across the curriculum" and therefore include topics in science, the social sciences and history, among others.

The student background questions included asking about what students read, both in and out of school; how often students read various kinds of materials; how often students read for enjoyment; use of the library; understanding the value of reading; and the reading behavior of people in the students' homes. Details of the non-cognitive assessment are included in Chapter 6.

The Year 15 assessment and accompanying background questions address the following issues:

- Has students' overall reading performance changed over the last 13 years? Over the last 9 or 4 years?
- Have patterns of reading performance changed over the same periods? Do these patterns vary for different groups of students? Do these patterns vary among students who report different reading activities or preferences?
- Does reading performance vary with different kinds of reading material--that drawn from particular subject areas, perhaps, or that which presents a particular reading task?
- Has students' ability to answer questions about reference materials and study skills topics changed over the past 13 years? Over the past 9 to 4 years?
- Has students' ability to write about what they have read changed since Year 11? If so, in what ways?
- Have students' evaluations of what they read changed since Year 11?

3.2.5 Reading Exercise Development Consultants*

Ms. Virginia Allery
Apple Valley, MN

Dr. Fernie Baca
School of Education
University of Colorado
Denver, CO

Ms. Sharon Branscome
Sevierville, TN

Ms. Opaline Brice
Englewood, CA

Dr. Robin Butterfield
Northwest Regional Educational
Laboratory
Portland, OR

Dr. Carita Chapman
Swift Elementary School
Chicago, IL

Dr. John Chapman
Michigan Department of Education
Haslett, MI

---

* See Reading Objectives, 1983-84 Assessment (1984) for a list of consultants who participated in developing reading objectives.

Ms. Avon Chrismore
Center for the Study of Reading
Champaign, IL

Ms. Nancy Ciarleglio
New Haven, CT

Dr. James Connor
Science Education Program
New York University
New York, NY

Ms. Virginia Cornue
National Organization of Women
New York, NY

James Cunningham
University of North Carolina
Chapel Hill, NC

Dr. Billie Day
Teacher
Benjamin Banneker Academic High School
Washington, DC

Dr. Phil DiStefano
School of Education
University of Colorado
Boulder, CO

Ms. Margaret Gallagher
Center for the Study of Reading
Champaign, IL

Ms. Tee Gallay
Chicago, IL

Mr. Pete Garcia
Dixon, NM

Dr. Geneva Gay
Purdue University
West Lafayette, IN

Dr. Sandra Gibbs
National Council of Teachers of English
Urbana, IL

Ms. Carol Gibson
National Urban League
New York, NY

Mr. Gene Goff, Jr.
Poca, WA

Roseann Gonzalez
University of Arizona
Tucson, AZ

Dr. Kris Gutierrez
Director of Academic Affairs
University of Colorado

Ms. Carol Harner
Littleton, CO

Dr. Shirley Munoz-Hernandez
Columbia University
New York, NY

Mr. Jack Holmquist
York, NE

Dr. Shu-In Huang
City of Thornton
Thornton, CO

Peter Johnston
SUNY at Albany
Albany, NY

Dr. Henry B. Maloney
Teacher
Seaholm High School
Birmingham, MI

Carole L. Mathews
Boulder Valley Schools
Boulder, CO

Greg Morris
Pittsburgh Public Schools
Pittsburgh, PA

Ms. Rosa Casarez-Najera
Stanford, CA

Taffy Raphael
Michigan State University
East Lansing, MI

Dr. Linda Reed
Lakewood, CO

James Robinson
Boulder Valley School District
Boulder, CO

Dr. Mary Budd Rowe
College of Education
University of Florida
Gainesville, FL

Dr. Peter Sanders
College of Education
Wayne State University
Detroit, MI

Ms. Dorothy Sibley
Miami Chapter ASUW
Miami, FL

Ms. Lucille Stillwell
Bernalillo, NM

Dr. Violet Strahler
Dayton Public Schools
Dayton, OH

Dorothy Strickland
Columbia University
New York, NY

Dr. Barbara Swaby
School of Education
University of Colorado
Colorado Springs, CO

Barbara Taylor
University of Minnesota
Minneapolis, MN

Dr. Robert Tierney
Center for the Study of Reading
University of Illinois of
Urbana/Champaign
Champaign, IL

Celeste Woodley
Boulder Valley School District
Boulder, CO

Ms. Kathy Yen
San Francisco Public Schools
San Francisco, CA

Chapter 4

SAMPLE SELECTION AND INSTRUMENT COLLECTION

Morris H. Hansen
Benjamin J. Tepping
Josefina A. Lago
John Burke

Westat, Inc.

The sample design for the Year 15 NAEP generally follows earlier designs but introduces some changes to serve new goals and increase efficiency. One innovation makes it possible to provide estimates for the modal grades corresponding to ages 9, 13, and 17. Another is the introduction of a balanced incomplete block design combined with a spiralled procedure for assigning tests to students. This change serves important analytical purposes, reduces sampling error, and facilitates administration. A third design innovation includes a sample of teachers of sampled students to correlate teacher and student characteristics.

The sample for the Year 15 NAEP was a multi-stage probability sample, with counties or groups of counties serving as first-stage sampling units, elementary and secondary schools serving as second-stage sampling units, the assignment of sessions by type to sampled schools serving as a third stage of sampling, and the selection of students within schools and their assignment to sessions serving as the fourth stage of sampling.

A total of 64 first-stage units was included in the sample, and assessments were conducted at 1,465 schools. Various blocks or packages of exercises were administered in these schools to a total of about 30,000 students in each of the three grade/ages.

To facilitate the transition to a new organization (the Educational Testing Service [ETS] was the new grantee responsible for the NAEP project with Westat as the survey subcontractor) the sample of PSUs and schools was drawn by the Research Triangle Institute (RTI), the earlier survey subcontractor. These samples were drawn following the principles and methods developed by RTI, and similar to those of recent earlier assessments.[1] Procedures more or less similar to those of prior assessments were used for subsequent stages of sampling as well, but were modified to accommodate new goals adopted by ETS.

---

[1] See Final Report on National Assessment of Educational Progress: Sampling, Weighting, and Quality Check Activities for Assessment Year 13. June 1983 (RTI/1967/00-02F).

The principal new goals included the following:

(a) In earlier assessments, the students sampled and assessed were those in ages 9, 13, and 17. In the Year 15 assessment the decision was made to draw samples to assess students of ages 9, 13 and 17[2], and students of the corresponding modal grades 4, 8, and 11.

(b) In earlier assessments, test items had been assembled into various packages. The same package of items was administered to all students in a session, which usually consisted of a sample of about 20 students. In Year 15, ETS specified and developed a new procedure in which exercises were grouped into a larger number of smaller blocks, and assembled into test booklets in a balanced incomplete block (BIB) design. These booklets were then assigned to students in a rotating or "spiral" design so that different booklets were assigned to each student in a session.

In addition, some of the assessments were administered as in earlier assessments, to provide comparable procedures for measuring change. In these, all students were administered the same "package" of items, and in these sessions the questions were presented orally from a recorded tape as well as visually, or were paced by a tape recording.

(c) A questionnaire was obtained for a sample of teachers of sampled students, to permit correlation of teacher and student characteristics.

(d) Earlier assessments had identified and excluded from the assessment students with limited English proficiency or certain handicaps. For Year 15 such students were again excluded, but a questionnaire was obtained for a sample of them to allow additional description and analysis.

---

[2] The following birthdate ranges, consistent with previous assessments, were used to define the Year 15 age groups: January-December 1974 for Age 9; January-December 1970 for Age 13; and October 1966-September 1967 for Age 17. To maintain comparability with previous assessments, students of each age group, along with students in the corresponding modal grade, were assessed at the same times of year as in prior assessments. Times of assessment were: October-December for Grade 8/Age 13; January-February for Grade 4/Age 9; and March-May for Grade 11/Age 17.

Some other changes were made in an effort to reduce costs, or reduce sampling variances or nonresponse biases, or both:

(e) Assessments were administered in moderately larger session sizes for Year 15 than in earlier assessments.

(f) Adjustments for nonresponse were made session by session, as in the past, for the comparably administered taped assessments. Somewhat different adjustments for nonresponse were made for the assessments administered by the new spiral procedures.

(g) A post-stratification procedure was introduced to replace the earlier "smoothing" procedure.

A brief general description of the Year 15 survey design follows, including some discussion of the new features.

4.1 The Sample of First-Stage Units

The first-stage sample was a stratified sample of 64 primary sampling units (PSUs), drawn by RTI to represent the 50 states and the District of Columbia. Each PSU consisted of a county or a group of counties. Counties were grouped only as needed to achieve a specified minimum size in terms of numbers of eligible students. The number of PSUs to be selected for the sample and their minimum size were specified by Westat. The specified total of 64 PSUs to be selected was the same as the number used for the Year 13 assessment, and was deemed minimal but sufficient to control the PSU contribution to variance to a reasonable level. Following is a brief description of procedures followed by RTI for defining, stratifying and selecting the sample of PSUs[3].

(a) Twenty primary strata of counties were defined, using 1980 Census data, based on four geographic regions by five "Sample Description of Community" (SDOC) classes. The latter separately identified (1) SMSA counties containing at least 10,000 or more population in a big city (a city of 200,000 population or more), (2) remaining counties in "big city" SMSAs, (3) other counties containing any part of a city of 25,000 or more population, (4) all other counties not identified as extreme rural, and (5) counties identified as extreme rural (i.e., not having 10,000 or more urban population, non-zero farm employment, and classified as extreme rural on the basis of an occupational index).

---

[3] For a detailed description of the selection of PSUs, see the RTI Final Report (RTI/2589/03-00F), Primary Sample for Years 15-19 of the National Assessment of Educational Progress.

(b) Preliminary measures of size were computed for each county (frame unit) by separately estimating the enrollment of 9-, 13-, and 17-year-old students in elementary and secondary schools for each county, using Quality Education Data, Inc. (QED)[4] data on school grade-range and total enrollment, and using prediction formulas developed by RTI on the basis of prior experience. The preliminary measure of size was the average enrollment of the three age classes.

(c) Adjusted measures of size were computed by doubling the preliminary measures of size for counties identified as extreme rural and for low socio-economic status (Low-SES) tracts of "big" cities. (Low-SES Census tracts were identified within the central big cities in the counties included in SDOC class 1, based on an index of SES computed for each Census tract.)

(d) The number of PSUs to be sampled was allocated to the 20 primary strata, approximately in proportion to the adjusted measures of size.

(e) Within the 20 primary strata, PSUs consisting of one or more counties were defined within states (with minor exceptions), each PSU to include a minimum adjusted measure of size of 1,000. The PSUs within each primary stratum were then ordered by state (after states within a region were ordered in a serpentine manner), and by percent minority within state (with reverse ordering in successive states).

(f) PSUs were selected by a sequential zone selection algorithm developed by Chromy (1979). For small PSUs (i.e., those with adjusted measures of size smaller than the zoning interval), the selections were made without replacement, and with probability of selection proportional to the adjusted measures of size. For such PSUs the use of the algorithm made the PSU sampling weight inversely proportional to the adjusted measure of size of the PSU. Larger PSUs could be selected more than once; in fact, two large PSUs were each selected twice.

---

[4] Quality Education Data, Inc. (QED) maintains and updates annually lists of schools showing grade span, total enrollment, school district, principal's name and other information for each school. The initial data provided by QED were evaluated against Census school-enrollment data by RTI, which led to some corrections of the QED file, made before the data were used in computing measures of size for sampling.

4.2 The Initial Sample of Schools

An initial sample of 1,682 schools was selected from the 64 primary sampling units, with the selections carried out independently for the three age classes. A total of 700 schools was selected for Age 9 (and Grade 4), 588 for Age 13 (and Grade 8), and 394 for Age 17 (and Grade 11)[5]. However, some schools contained eligibles for two or more of the age classes and were selected more than once so that a total of 1,587 distinct schools was selected. Enough schools were selected within an age class in each PSU to yield the desired sample size of students, with a reserve to allow for some ineligible schools and for some non-participation of schools, based on Year 13 experience.

Often, a relatively efficient procedure is to draw the sample with varying probabilities at the various stages of sampling, but such that the overall probability of selection of a final unit in the sample (in this case a student selected to take a particular type of assessment booklet) is the same for each student. With some exceptions in which oversampling or undersampling was done by design, this was a goal in the NAEP sample, and it affected design decisions for sampling PSUs and schools as well as later stages of sampling.

To control costs, the sample of schools was selected to allow a maximum of about 200 age or grade eligibles to be invited to assessment sessions in a school for Grade 4/Age 9 and up to about 250 age or grade eligibles for Grade 8/Age 13 and Grade 11/Age 17. While these specifications allow relatively large samples of students from some individual schools, the average number of students assessed per school was well below the maximum. Moreover, only a small fraction of students assessed in a school is assessed for a given block of exercises. It was recognized that variances would be increased by allowing maximum cluster sizes up to these levels but perhaps not unduly in relation to costs.

After initial study, it was estimated that the number of students in a school that were eligible by either age or modal grade would average roughly 1.3 times the number of age eligibles. This would vary by age class and from school to school, but not widely; the number of age eligibles in a school would still provide satisfactory measures of size for use in sample selection.

As described below, varying but roughly equal measures of size were assigned to those schools containing estimated age-eligible students ranging from 20 to 160 (for Age 9) or to 200 (for Ages 13 and 17). Schools with less than 20 estimated age eligibles were selected with lower probabilities, and schools above the indicated maximum size were selected with probabilities proportional to the estimated numbers of age-eligible students (with approximately constant numbers of students to be subsampled from them).

---

[5] Three schools were selected twice for Age 17 and Grade 11.

With the adoption of these general specifications, the sampling of schools by RTI proceeded approximately as follows:

(a) The estimated number of age eligibles, E_i, was computed for school i, using QED information for school year 1982-83. The number in each grade was estimated by dividing total enrollment by the number of grades; the number of age eligibles was estimated by applying the RTI prediction formulas.[6]

(b) For the "big-city" PSUs

(i) An SES index was assigned to each school (based on employment, unemployment, occupational, and income data from the 1980 Census for each Census tract, and by approximately matching the zip codes to the Census tracts).

(ii) Schools were classified as Low-SES (Stratum 1), and Other (Stratum 2). After establishing a cutoff for the SES index to define the two strata, the schools were ordered by estimated number of age eligibles in ascending order in Stratum 1 and descending order in Stratum 2. For other PSUs the schools were ordered by size.

(c) A preliminary measure of size, s'_i was assigned to each school, based on the estimated number of age eligibles E_i, illustrated as follows for Age 9, for which $\bar{n} = 20$ is the planned full-session size:

(i) If school i had six or fewer estimated age eligibles, $s'_i = .25$;

(ii) If school i had seven to nineteen estimated age eligibles,

$$s'_i = E_i/20;$$

(iii) If school i had 20 or more age eligibles but less than 160,

$$s'_i = \frac{E_i}{20k_i}$$

---

[6] See Section 3.1.4 of School Sampling Procedure for Year 15 of the National Assessment of Educational Progress, September 1983 (RTI/2589/02-00F).

where k_i is the number of sessions of 20 that can be accommodated by E_i; and

(iv) If school i had 160 or more age eligibles

$$s'_i = \frac{E_i}{160} .$$

(d) A final measure of size, s_i, was computed for each school by doubling the preliminary measure of size for those schools in "big-city" PSUs that had been assigned to the low-SES stratum, and by using $s_i = s'_i$ for all other schools.

(Note that the extreme rural PSUs were already oversampled by a factor of 2, which had the effect of doubling the school sample in these.)

(e) The number of schools to be selected in an age class was computed separately for each PSU to yield approximately the desired number of students to be tested, after making approximate allowance for school and student nonresponse and for ineligible schools. The number of schools to be selected, t, is

$$t = \frac{\bar{n}m}{\bar{k}}$$

where

$\bar{n}$ is the number of students per full age session (e.g., 20 for Age 9);

m is the number of full age-eligible sessions assigned to the PSU;

$$\bar{k} = \frac{\Sigma s'_i k_i}{\Sigma s'_i}$$

that is, the weighted average of the k_i (the number of age-eligible sessions available in school i, as used in computing the measures of size); and

s'_i is defined above.

(f) The t schools were then selected in the PSU for the age class by sampling with probabilities proportionate to the measures of size, s_i. It was recognized that a school might be selected twice for the same age class by this procedure, and

> thus (to avoid administering more than ten sessions in a school) it might be necessary to transfer sessions to another sampled school. (Actually, only three schools were selected twice, and these were for Age 17.)

A detailed description of the initial selection of the sample schools is given in the RTI final report cited previously.

4.3 Updating the School Sample

ETS made the initial contacts with sampled school districts to obtain participation. The districts were then requested by Westat to identify schools that were new since the time of the QED list, or schools with changes in grade range or major changes in enrollment. These were given appropriate chances to be in the sample using probability-sampling procedures. Also, the sample was supplemented in a few PSUs where losses due to closed schools or other changes left too few schools in the sample. A Principal Questionnaire showing updated grade and enrollment figures and certain other school characteristics was requested from each of the cooperating schools prior to the assessment.

Some substitutions were made, as needed and to the extent feasible, for non-cooperating schools. Generally, substitutions were made for schools refusing to participate in the assessments if their omissions would result in an unacceptable balance in school type among the schools assessed, according to the size of the school and the socio-economic status of the community, or would result in a substantial reduction in the number of students tested. In general, substitution of schools was made within the same PSU, but in a few cases losses in one PSU were compensated for by additional assessments in the sampled schools in another PSU. In three cases substitute schools were obtained from a neighboring and similar county (not a member of the primary sample of PSUs).

Table 4(1) summarizes the selection and participation of schools. The cooperation rates obtained were approximately the same those as obtained for the Year 13 NAEP (an overall rate of 88.1 for Year 15 and of 88.0 for Year 13).

4.4 The Assignment of Sessions to Schools, by Type

The assignment of sessions to schools was done separately by the two types of sessions, designated "spiral" and "tape."

As discussed in Chapter 5, the balanced incomplete block (BIB) design together with spiralling (or interspersing) the assessment booklets was introduced for the first time in Year 15. This made it possible to correlate results for all pairs of exercises in the BIB design. The exercises were divided into blocks of items, each block also containing some background questions. The blocks were assembled into 63 test

Table 4(1)
Summary of NAEP Year 15 School Participation Experience

	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17	Total Sample
Initially selected schools	700	588	394	1,682
Supplemental selections	17	2	1	20
New schools added to sample	2	1		3
Total original sample	719	591	395	1,705
Out-of-range or closed (A)	15	12	17	44
No eligibles enrolled (B)	17	64	17	98
District refused (C)	61	42	40	143
School refused (D)	19	14	21	54
Cooperating - No student sample (F)	0	4	1	5
Cooperating - Assessment conducted (E)	607	455	299	1,361
Cooperation rate = $\frac{B+E+F}{B+C+D+E+F}$	88.6	90.3	83.9	88.1
(Year 13)	(88.0)	(89.2)	(86.5)	(88.0)
Replacement for refusals*	67	28	34	129
Out-of-range or closed	3	0	0	3
No eligibles enrolled	5	3	1	9
Refusals	5	2	6	13
Assessment conducted	54	23	27	104
Total contacted schools	786	619	429	1,834
Total assessments conducted	661	478	326	1,465

*Includes schools added through the partial PSU replacement procedure and school-by-school substitution booklets, most containing three blocks as well as a set of background questions common to all the booklets, so that each block occurred in the same number of booklets and each pair of blocks occurred in the same number of booklets. As a result, it was expected that each block of items would be administered to about 2,000 students in each grade/age and each pair of blocks would be administered to about 200 students in each grade/age. The booklets were assembled systematically into packages, arranged so that the starting booklet varied from session to session.

The tape design used an administration procedure like that of earlier NAEP assessments so as to provide direct comparison with the results of earlier assessments and to calibrate the results of the spiral design. The administration of each booklet used a tape recording, as in earlier assessments. The specified sample size was such that each tape-administered booklet was expected to be administered to about 1,250 students.

A preliminary allocation of sessions was made to the sampled schools based on the QED 1982-83 information on enrollment and grade range for use in making initial arrangements with the schools. These were revised later on the basis of the Principal Questionnaire which provided enrollment by grade and information on SES status and minority enrollment for the school.

For the purpose of this allocation, small schools were clustered with others in the sample so that there was an estimated minimum of eight (and usually more) age-eligible students in each school cluster. The allocation of tape sessions was made first, by ordering the school clusters by an index of socio-economic status (based on the information provided in the Principal Questionnaire) and by size, then selecting a systematic sample of four school clusters with probability approximately proportional to the estimated number of age-eligible students in the school cluster. The next step was to assign one spiral session to each school cluster not selected for a tape session and to allocate the balance of the spiral sessions specified for the PSU to school clusters approximately proportionate to the estimated number of students (eligible by age or grade) that would be available after the initial assignment of tape and spiral sessions. Details of the allocation appear in the Report on Sample Selection, Weighting and Variance Estimation: NAEP--Year 15 (Lago, Burke, Tepping, & Hansen, 1985).

4.5 The Samples of Students

A total of about 29,300 students was to be tested for each grade/age, including students for the corresponding modal grade. This means an average of about 460 completed assessments per PSU for each grade/age. On the basis of the experience in Year 13, conservative estimates were made of the proportion of students that would be excluded from testing because of language or other disability and of the proportion of students invited for assessment that would actually complete the assigned test. These estimates led to the determination of the sampling rate to be applied in each sample school. Since the estimates were conservative, the number of students assessed was expected to exceed the target. For Grade 4/Age 9, 31,579 students were assessed; for Grade 8/Age 13, 33,563 students were assessed; and for Grade 11/Age 17, 35,070 students were assessed.

A Student Listing Form (SLF) was filled out for each participating school; all enrolled students of the specified age (9, 13 or 17) and all others in the corresponding modal grade (4, 8 or 11) were to be entered on the SLF in any order convenient for the school. In a few instances for very large schools, only a sample of students was listed on the SLF. The SLF was ordinarily prepared by the school, but Westat staff assisted or prepared the form when desirable or necessary.

After the SLF was completed the selection of sample students was carried out briefly as follows:

(a) A computer generated listing of sample SLF line numbers was prepared in advance by Westat to identify the students to be included in the sample. When the number of students listed on the SLF differed widely from the anticipated number, communication was handled by telephone and a new set of sample line numbers was supplied.

(b) The sample line numbers also identified the type of session, spiral or tape, to which a sampled student was assigned.

(c) The names of students selected for the sample were reviewed by appropriate school personnel to identify sampled students who for language reasons or certain types of handicaps would be unable to take the test and thus should be excluded.

Makeup sessions were scheduled in schools in which the students assessed constituted less than 75 percent of the selected sample in the case of spiral sessions, less than 50 percent in the case of tape sessions for 9-year-olds and 13-year-olds, and less than 75 percent in the case of 17-year-olds. Very few makeup sessions were necessary for 9- and 13-year-olds. For the 17-year-olds, makeup sessions were conducted in about 20 percent of the sample schools.

4.6 The Sample of Excluded Students

The Year 15 assessment, as in previous assessments, excluded students who were functionally handicapped to the extent that they could not participate in the assessment as it was normally conducted. Specific groups excluded were:

(1) students with limited English proficiency;

(2) students identified as having behavioral disorders; and

(3) students physically or mentally handicapped, including Educable Mentally Retarded (EMR), in such a way that they

could not respond to NAEP exercises as they were normally administered.

In Year 15 a sample of excluded students was drawn and data collected about them. In most cases, students to be excluded from assessment were identified before sampling but were sampled at the same rates as any other eligible student. In other cases, excluded students were identified only for students selected for the sample.

For each sampled excluded student, an Excluded Student Questionnaire, which focused on the nature of the student's problem and the school's approach to handling it, was filled out by school personnel. This data collection effort for excluded students was a new feature of the Year 15 assessment permitting national estimates of this subgroup of age- and grade-eligible students. Table 4(2) shows the distribution of excluded students by reason for exclusion for the three grade/ages.

4.7 Student Participation Results

The NAEP sample was designed to yield a target number of spiral assessment and of each of the four tape assessments. Table 4(3) compares the target assessments to the actual assessments for the three grade/ages.

As indicated previously, the allocation of sessions to schools and sampling rates within schools were based on the Year 13 proportion of excluded students identified and student participation rate, and the Year 15 target number of completed assessments. Tables 4(4) and 4(5) compare the Year 13 and Year 15 proportion of excluded students and student participation rates, respectively. As shown, the student participation rates in Year 15 were about 2 percent higher (for Grade 4/Age 9 and Grade 8/Age 13) and 8 percent higher (for Grade 11/Age 17) than the participation rates for corresponding age classes in Year 13. Also, the losses due to excluded students were smaller for Grade 4/Age 9 and Grade 8/Age 13 in Year 15. As a result, and because some reserves were provided for in allocating the sample to allow for the possibility of greater losses than anticipated on the basis of Year 13 experience, the Year 15 actual assessments shown in Table 4(5) were considerably higher than the target of 29,267 assessments per grade/age.

4.8 The Associated Teacher-Student Sample

In addition to the student data collection effort, NAEP also collected data on a sample of English or language arts teachers who were identified as the principal such teacher of a subsample of one or more of the grade/age-eligible students in the spiral sample. The objective of the survey was to collect for analysis data that involve the characteristics of a student's teacher.

The teachers who participated in the teacher survey were selected as follows: From those students selected for spiral sessions in a school, a subsample of students was selected equal to the number of spiral sessions

Table 4(2)

Weighted and Unweighted Distribution of Excluded Students, by Reason for Exclusion and Grade/Age

Grade 4/Age 9

	Unweighted		Weighted	
Reason	Count	Percent	Count	Percent
A Physical or mental handicap	761	54	91,538	53
B Behavioral disorder	102	7	11,488	7
C Handicap and limited English proficiency	102	7	11,488	7
D Limited proficiency in English	453	32	56,922	33
All reasons	1,418*	100	171,436	100

Grade 8/Age 13

	Unweighted		Weighted	
Reason	Count	Percent	Count	Percent
A Physical or mental handicap	971	67	120,261	67
B Behavioral disorder	102	7	13,117	7
C Handicap and limited English proficiency	86	6	12,440	7
D Limited proficiency in English	289	20	33,236	19
All reasons	1,448	100	179,054	100

Grade 11/Age 17

	Unweighted		Weighted	
Reason	Count	Percent	Count	Percent
A Physical or mental handicap	817	59	68,042	59
B Behavioral disorder	48	4	4,733	4
C Handicap and limited English proficiency	106	8	8,824	8
D Limited proficiency in English	390	29	33,563	29
All reasons	1,361	100	115,162	100

*Two Grade 4/Age 9 excluded students were not retained on the NAEP database due to insufficient data.

Table 4(3)

Comparison of Year 15 Target Assessments to Actual Assessments, by Grade/Age

	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
	Target	Actual	Target	Actual	Target	Actual
Spiral assessments	24,267	26,087	24,267	28,405	24,267	28,861
Tape assessments*	5,000	5,492	5,000	5,158	5,000	6,209
Booklet 64	1,250	1,403	1,250	1,310	1,250	1,539
Booklet 65	1,250	1,356	1,250	1,276	1,250	1,540
Booklet 66	1,250	1,389	1,250	1,283	1,250	1,596
Booklet 67	1,250	1,344	1,250	1,289	1,250	1,534
Total	29,267	31,579	29,267	33,563	29,267	35,070

* Tape assessments were administered to age only.

Table 4(4)

Comparison of Year 13 and Year 15 Proportion of Excluded Students, by Grade/Age

Grade/Age	Year 13 Excluded (%)*	Year 15 Excluded (%)
Grade 4/Age 9	5.1	4.3
Grade 8/Age 13	5.2	4.1
Grade 11/Age 17	3.5	3.7

* Year 13 assessment was administered to age only.

Table 4(5)

Comparison of Year 15 and Year 13 Student Participation Rates, by Type of PSU and Grade/Age

	Assessed (a)	Assessed (b)	Invited to Assessment (c=a+b)	Participation Rate (a/c)	Year 13 Participation Rate**
Grade 4/Age 9					
PSU Type A*	22,101	2,336	24,437	90.4	90.5
PSU Type B*	9,478	694	10,172	93.2	90.5
Total	31,579	3,030	34,609	91.3	90.5
Grade 8/Age 13					
PSU Type A*	23,234	3,563	26,797	86.7	85.0
PSU Type B*	10,329	1,342	11,671	88.5	90.0
Total	33,563	4,905	38,468	87.3	85.5
Grade 11/Age 17					
PSU Type A*	25,406	5,700	31,106	81.7	66.0
PSU Type B*	9,664	1,592	11,256	85.9	82.0
Total	35,070	7,292	42,362	82.8	74.2

* PSUs Type A are the urban PSUs (SDOCs 1, 2 and 3); PSUs Type B are the non-urban PSUs (SDOCs 4, 5 and 6).

**Year 13 assessment was administered to age only.

assigned to the school. The principal English or language arts teacher for each of these sample students was identified by the school and was asked to complete a Teacher Questionnaire. Before an assessment began, all students in each session were asked to code their principal English teacher in the box provided on the cover of the exercise booklet. Thus, it was possible to associate the sample of teachers with assessed students.

The conditional probability that a spiral assessment selected student (of teacher k) had his or her teacher in the survey is given by

$$P_k = 1 - \frac{\binom{n-n_k}{t}}{\binom{n}{t}}$$

where the symbol $\binom{a}{b}$ denotes the number of combinations of a things taken b at a time, and

n = the total number of students invited to spiral assessments;

n_k = the number of students invited to spiral assessments whose teacher is the k^{th} teacher of the school, k=1, 2, ..., K; and

t = the number of spiral-invited students subsampled for the teacher survey.

If any English teacher was identified for more than one of the subsampled students, the teacher completed only one questionnaire. Thus, the number of completed questionnaires was smaller than the number of students subsampled for the teacher survey.

Since the principal teacher was recorded only for assessed students, P_k was approximated by replacing n_k and n by the numbers of assessed rather than invited students. Students whose teachers were surveyed have their weights multiplied by the reciprocal of P_k in any analyses that involve relating teacher characteristics to student characteristics. The weights were further adjusted, within PSUs, to account for the fact that not all assessed students indicated their principal language arts teacher and not all sampled teachers returned a completed questionnaire. They were also adjusted within PSUs by a post-stratification procedure so that the sum of the weights for students in the teacher sample were equal to the sum of the weights for all students in the spiral sample. From the figures shown in Table 4(6), because of either nonresponse or overlap we lost about 25 percent of the Grade 4/Age 9 sampled teachers, 38 percent of the Grade 8/ Age 13 sampled teachers and 35 percent of the Grade 11/Age 17 sampled teachers. In addition, about 2 percent of the teachers completing a questionnaire were not linked to an assessment booklet; that is, the subsampled student through whom the teacher was brought into the sample was either absent, excluded, or had recorded a different teacher than had been recorded by the school as the student's principal language arts teacher, and no other tested student had reported that teacher.

Table 4(6)

Distribution of Teachers by Grade/Age and Participation Status

Grade/Age	Sampled Teachers	Distinct Teachers Sampled	Responding Teachers	Teacher Response Rate	Linked Teachers
Grade 4/Age 9	1,361	1,066	1,025	96	1,004
Grade 8/Age 13	1,275	821	790	96	779
Grade 11/Age 17	1,406	980	915	93	901
Total	4,042	2,867	2,730	95	2,684

Chapter 5

THE ASSIGNMENT OF EXERCISES TO STUDENTS[1]

Albert E. Beaton
Eugene G. Johnson
John J. Ferris

Educational Testing Service

The purpose of the National Assessment of Educational Progress (NAEP) is to estimate the performance in particular subject areas of various subgroups of students, at specific age or grade levels. In the past as well as at the present, NAEP has aimed at providing information on a broad spectrum of appropriate and important skills and performances in the subject areas it has assessed. This information has been and continues to be provided at the level of subgroups of the population rather than at the level of the individual student.

To accomplish this purpose, there is no need for precise measures for any individual student. Consequently, it is not necessary or even desirable that each individual student take the entire battery of exercises designated for the student's grade or age level. In addressing the problem of estimating the proportion of a population who could correctly respond to a population of items (given a fixed number of item responses), Lord (1962) has shown that a sample with many persons taking just one item each resulted in an estimator with a smaller standard error than one derived from a sample in which fewer persons responded to many items. Since such a sampling scheme is ordinarily not cost-effective because selecting individuals is expensive, a number of exercises are presented to each sampled individual.

Both ETS and ECS (the previous grantee for NAEP) employed multiple matrix sampling techniques for the assignment of a set of exercises to subsamples of students. The matrix sampling approaches in both cases enable broad coverage of a given subject area in terms of the total number of exercises which can be assessed while restricting the effort required of any individual student. The two approaches, however, have some fundamental differences.

The ECS multiple matrix sampling design divided the entire pool of exercises designated for a given age group into a number of distinct sets, called packages, each of which would take a student about three quarters of an hour to complete. Using this approach, the six hours of assessment exercises allocated to an age group would result in eight packages. Since no student was administered more than one package, this simple matrix design allowed the calculation of measures of relation between exercises within the same package but not between exercises in different packages.

---

[1] The tables and figures for this chapter were produced by David Freund.

To remedy this deficiency, ETS has chosen a complex variant of multiple matrix sampling called Balanced Incomplete Block (BIB) spiralling. This approach continues to allow the broad coverage of subject areas and also allows the study of the interrelationships among all exercises within and between subject areas. The basic idea is to divide up the total assessment time into small blocks. Each exercise block is then assigned to a number of assessment booklets such that each block of exercises is paired with each other block in some booklet. The booklets are then spiralled so that students in an assessment session are given different booklets. Using BIB spiralling, a large number of booklets must be created, but the interrelationships between objectives may be examined since each exercise is paired with each other exercise in some booklet.

The BIB spiralling method of exercise assignment has another advantage over the previous technique. In the ECS method of item administration, after the sample of students within a school was selected and brought to an assessment session, the same package of exercises was distributed to all students within that session. This administration of the same exercises to clusters of students within a school was necessary because the administration of a package was accompanied by a paced audiotape of the exercise stimuli, designed to minimize the effect of a student's reading ability on performance in other subject areas. Unfortunately, this administration of the same exercises to clusters of students within schools also results in a potential increase in sampling variability over a simple random sample of the same number of students because of intra-cluster correlation. In contrast, in the spiralled mode of item administration a set of exercises is presented to fewer persons in a school and to more schools. This results in a marked reduction in the intra-school cluster effect over the package administration procedure of previous assessments. Consequently, since the sample of students is more efficiently utilized, the required sample size to achieve a given standard error is reduced. Alternatively, the standard error for a given sample size will be reduced.

The remainder of this chapter will detail the spiralling process as implemented for the NAEP and will discuss its perceived advantages and disadvantages. First, however, the considerations in developing the design of the assessment instruments and the interplay between the amount of substantive coverage and the sample size will be discussed.

5.1 Considerations in the NAEP Assessment Design

The design of any study is circumscribed by the amount of funds available; thus, the NAEP staff had to decide how to allocate its resources to allow the broadest possible assessment of its Year 15 subject areas, reading and writing. The decisions that resulted in the final design were as follows:

(1) Each student would be asked to participate for about three quarters of an hour. To have a national assessment at all requires the cooperation of schools, and we felt, as did the ECS staff before us, that limiting the intrusion on individual students to about one class period would help us gain acceptance in the schools. The design originally called for 46 minutes of assessment for each student, but was extended to 48 minutes when a review of the early data showed that students were not reaching some important background and attitude questions.

(2) The available funds were sufficient to gather data on about 30,000 students at each grade/age level. It should be noted that, under the terms of the grant, the Research Triangle Institute provided the sample of schools. Westat, who is the ETS subcontractor for sampling and field administration, reviewed the sample and studied some preliminary data collection plans to estimate the number of students who could be assessed for the available funds. Thirty thousand students at 48 minutes per student resulted in an expected total of 24,000 hours of testing time at each grade/age level.

(3) Each exercise would be responded to by about 2,600 students at each grade/age level. In past assessments, around 2,500 to 2,600 students at each age level were targeted for each exercise. We felt that the efficiencies of spiralling would allow us to reduce the number of students from about 2,500 taking each exercise (as in earlier years) to about 2,000 without an increase in sampling error. However, we were committed to sample both the age levels which were sampled in the past (ages 9, 13, and 17) and the grades into which most of those youths fell (grades 4, 8, and 11). We estimated that a sample of 2,600 at each grade/age level would result in a sample of about 2,000 at each age and also about 2,000 at each grade.

(4) Five thousand students at each age level would be designated to receive audiotaped assessment. Data has been collected in national assessments since 1969, and we did not want to lose continuity with the data already collected. Since we were making a change from audiotaped administration to pencil-and-paper administration, we felt that we needed to determine what effect the method of administration had on the performance of students on assessment exercises. Therefore, a sample of 5,000 students was designated for assessment using the same procedures as in the past.

(5) There would be six minutes of questions common to all students. Some questions, such as racial/ethnic

identification, are so important in the assessment that they must be asked of every student. At first, four minutes were allowed for such questions, but early experience required us to increase this section to six minutes.

(6) Assessment exercises and other background and attitude questions would be grouped into blocks which would require fourteen minutes to complete. These blocks would contain an average of twelve minutes of reading and writing exercises and two minutes of background and attitude questions. Thus, each student's 48 minutes would include the common questions (six minutes) and three blocks of other assessment questions (fourteen minutes each). In terms of content, a student would spend twelve minutes on background and attitude questions and 36 minutes on reading or writing exercises.

(7) Several longer writing exercises could not be administered in the twelve minutes allocated in each block and were accommodated by creating three double-length blocks (28 minutes).

It is immediately clear that a perfectly balanced incomplete block design is impossible, since the double-length blocks can not be paired within the 48-minute time limit. Although we could not assign two double-length blocks to any student, we could assign them in such a way that we could compare the double-length blocks indirectly through one or a chain of single-length blocks, and we did.

The final sample consisted of three parts, one of which received BIB-spiralled booklets, a second received partially BIB-spiralled (UBIB) booklets, and the third was a matrix sample which was assessed using paced audiotapes. The target sample sizes and the amount of assessment time for the different samples are shown in Table 5(1).

5.2 The Balanced Incomplete Block (BIB) Spiral Sample

The booklets in the BIB design each contain the common block and three of the nineteen single-length blocks assigned to this sample. The nineteen blocks were assigned to booklets using a cyclic Youden rectangle (see Beall, 1971). This procedure required the formation and printing of 57 different booklets and assigned each individual block to precisely nine different booklets. Each block is combined with each other block exactly once in this design, and thus each pair of exercises was assigned to some sample of youths. (The block assignments are shown in the left half of Table 5(2).)

Block designations were re-coded using a permutation mapping of the nineteen letters A through T (except I--there is no block I). The booklet numbers were then re-coded using a permutation mapping of the integers 1 through 57. Finally, the block orders were randomly permuted within each

Table 5(1)

Sample Design Summary

Sample	----Blocks----		Booklets	-----Students per-----			-Assessment Time in Minutes-			
	Single	Double		Booklet	Block	Sample	Common	Subject Matter	Other	Total
BIB										
Age and Grade	19	0	57	156	1,400	8,867	6	228	38	272
Age Only				67	600	3,800				
Grade Only				67	600	3,800				
Total	19	0	57	290	2,600	16,467	6	228	38	272
UBIB										
Age and Grade	4*	3	6	700	1,400	4,200	6	120	20	146*
Age Only				300	600	1,800				
Grade Only				300	600	1,800				
Total	4*	3	6	1,300	2,600	7,800	6	120*	20	146*
TAPE (Age Only)	12	0	4	1,250	1,250	5,000	6	144	24	174
TOTAL-EACH GRADE/AGE	21	3	67			29,267	6	324	54	384
TOTAL-ALL GRADE/AGES	63	9	201			87,801			**	

* Two single blocks are duplicated in BIB sample
** Total assessment time depends on common blocks across grade/age

Table 5(2)

Booklet Design--BIB Spiral Sample
(19 x 3 x 57 Cyclic Youden Rectangle)

Original Design				Permuted Design			
	Item Block				Item Block		
Booklet	1	2	3	Booklet	1	2	3
1	A	B	G	1	T	G	L
2	B	C	H	2	A	L	P
3	C	D	J	3	D	A	T
4	D	E	K	4	C	S	E
5	E	F	L	5	C	A	H
6	F	G	M	6	G	F	H
7	G	H	N	7	K	R	N
8	H	J	O	8	R	M	F
9	J	K	P	9	O	N	L
10	K	L	Q	10	F	D	B
11	L	M	R	11	E	M	A
12	M	N	S	12	S	H	B
13	N	O	T	13	M	K	D
14	O	P	A	14	T	N	J
15	P	Q	B	15	M	T	C
16	Q	R	C	16	C	L	Q
17	R	S	D	17	H	E	R
18	S	T	E	18	C	P	F
19	T	A	F	19	L	S	K
20	A	C	L	20	N	B	E
21	B	D	M	21	N	C	D
22	C	E	N	22	Q	K	H
23	D	F	O	23	L	H	D
24	E	G	P	24	A	S	R
25	F	H	Q	25	L	J	R
26	G	J	R	26	T	F	Q
27	H	K	S	27	C	K	J
28	J	L	T	28	O	J	S
29	K	M	A	29	Q	O	D
30	L	N	B	30	B	Q	J
31	M	O	C	31	O	T	H
32	N	P	D	32	B	M	L
33	O	Q	E	33	C	R	O

Table 5(2)
(continued)

Booklet Design--BIB Spiral Sample
(19 x 3 x 57 Cyclic Youden Rectangle)

Original Design				Permuted Design			
	Item Block				Item Block		
Booklet	1	2	3	Booklet	1	2	3
34	P	R	F	34	G	O	E
35	Q	S	G	35	S	Q	M
36	R	T	H	36	B	A	O
37	S	A	J	37	K	G	A
38	T	B	K	38	O	F	K
39	A	D	H	39	P	S	T
40	B	E	J	40	E	F	L
41	C	F	K	41	H	M	J
42	D	G	L	42	J	E	D
43	E	H	M	43	F	J	A
44	F	J	N	44	B	G	C
45	G	K	O	45	P	B	K
46	H	L	P	46	S	F	N
47	J	M	Q	47	P	Q	E
48	K	N	R	48	B	R	T
49	L	O	S	49	P	M	O
50	M	P	T	50	R	P	D
51	N	Q	A	51	G	R	Q
52	O	R	B	52	S	G	D
53	P	S	C	53	H	P	N
54	Q	T	D	54	T	E	K
55	R	A	E	55	M	G	N
56	S	B	F	56	A	N	Q
57	T	C	G	57	G	J	P

booklet. The final design is shown in the right half of Table 5(2). The booklets in which each block appeared are listed in Table 5(3).

As shown in Table 5(1), this design called for each booklet to be administered to 288.9 different students and, since each block was in nine booklets, each block was therefore to be given to about 2,600 students, our target, at each grade/age combination. Altogether, this part of the design called for 288.9 students to take one of 57 booklets and thus 16,467 students in all. Looking at the age and grade samples separately, we expected each booklet to be administered to 222.2 youths at each age or grade level, thus each block to be administered to 2,000 youths, resulting in a total age or grade sample of about 12,667.

5.3 The Unbalanced Incomplete Block (UBIB) Spiral Sample

The booklets in the unbalanced design each contain the common block, a single-length block, and a double-length block. This design used seven blocks: three double-length blocks, two "new" single-length blocks that were not used in the completely balanced design, and two "old" blocks that were also used in the other design. This design resulted in the formation and printing of six booklets. Two of the double-length blocks were combined with one of the new and one of the old blocks; the other double-length block was paired with both of the new blocks. The assignment of blocks to booklets is shown in Table 5(4).

The design called for each of these booklets to be administered to 1,300 youths and, since each of the new blocks was in exactly two booklets, each block was also administered to 2,600 youths. Altogether, the design called for 7,800 students to take a UBIB booklet. The design also met the objective of having about 2,000 students take each exercise if we observed the sample for a particular age, or 2,000 students if we observed a particular grade.

The two booklets which contain a double-length block and one of the single-length blocks from the completely balanced sample result in an oversampling of these two single-length blocks since they are already adequately sampled in the BIB design. These two single-length blocks occur in nine BIB booklets, each of which is administered to about 289 students, and in one UBIB booklet, which is administered to 1,300 youths; thus, the targeted sample for each of these blocks was 3,900.

5.4 Overall Pairings of Item Blocks in the BIB/UBIB Design

The number of pairings of item blocks for all BIB and UBIB blocks are shown in Table 5(5). Because block Q replaced block Y for Grade 4/Age 9, the pairings for that grade/age sample are slightly different.

Table 5(3)

Spiral Sample
Block-to-Booklet Correspondence

Block	Booklet Numbers										
A	2	3	5	11	24	36	37	43	56		
B	10	12	20	30	32	36	44	45	48		
C	4	5	15	16	18	21	27	33	44		
D	3	10	13	21	23	29	42	50	52		
E	4	11	17	20	34	40	42	47	54		
F	6	8	10	18	26	38	40	43	46		
G	1	6	34	37	44	51	52	55	57		
H	5	6	12	17	22	23	31	41	53		
J	14	25	27	28	30	41	42	43	57	58	
K	7	13	19	22	27	37	38	45	54		
L	1	2	9	16	19	23	25	32	40		
M	8	11	13	15	32	35	41	49	55		
N	7	9	14	20	21	46	53	55	56		
O	9	28	29	31	33	34	36	38	49		
P	2	18	39	45	47	49	50	53	57		
Q	16	22	26	29	30	35	47	51	56	59*	63*
R	7	8	17	24	25	33	48	50	51	60	
S	4	12	19	24	28	35	39	46	52		
T	1	3	14	15	26	31	39	48	54		
U	58	59									
V	60	61									
W	62	63									
X	61	62									
Y	59**	63**									

* Grade 4/Age 9 only
** Grade 8/Age 13 and Grade 11/Age 17 only

Table 5(4)

Booklet Design
UBIB Spiral Sample

Booklet	Long Block	Short Block
58	U	J
59*	U	Y
60	V	R
61	V	X
62	W	X
63*	W	Y

* In Grade 4/Age 9, Block Q was substituted for Block Y in Booklets 59 and 63

Table 5(5)

Number of Pairings of Item Blocks in Spiral Design
(Number of Block Occurrences on the Diagonal)

Grade 4/Age 9

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
A	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
B		9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
C			9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
D				9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
E					9	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
F						9	1	1	1	1	1	1	1	1	1	1	1	1	1				
G							9	1	1	1	1	1	1	1	1	1	1	1	1				
H								9	1	1	1	1	1	1	1	1	1	1	1				
J									10	1	1	1	1	1	1	1	1	1	1	1			
K										9	1	1	1	1	1	1	1	1	1				
L											9	1	1	1	1	1	1	1	1				
M												9	1	1	1	1	1	1	1				
N													9	1	1	1	1	1	1				
O														9	1	1	1	1	1				
P															9	1	1	1	1				
Q																11	1	1	1	1		1	
R																	10	1	1		1		
S																		9	1				
T																			9				
U																				2			
V																					2		1
W																						2	1
X																							2

Table 5(5)
(continued)

Number of Pairings of Item Blocks in Spiral Design
(Number of Block Occurrences on the Diagonal)

Grade 8/Age 13 and Grade 11/Age 17

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
A	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
B		9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
C			9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
D				9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
E					9	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
F						9	1	1	1	1	1	1	1	1	1	1	1	1	1					
G							9	1	1	1	1	1	1	1	1	1	1	1	1					
H								9	1	1	1	1	1	1	1	1	1	1	1					
J									10	1	1	1	1	1	1	1	1	1	1	1				
K										9	1	1	1	1	1	1	1	1	1					
L											9	1	1	1	1	1	1	1	1					
M												9	1	1	1	1	1	1	1					
N													9	1	1	1	1	1	1					
O														9	1	1	1	1	1					
P															9	1	1	1	1					
Q																9	1	1	1					
R																	10	1	1		1			
S																		9	1					
T																			9					
U																				2				1
V																					2		1	
W																						2	1	1
X																							2	
Y																								2

5.5 Spiralling

The method for spiralling booklets was designed for two purposes:

(1) To achieve a ratio of nine students taking a UBIB booklet to two students taking a BIB booklet in order to meet the targeted sample sizes in each category; and,

(2) To distribute the booklets across the sample of students so that the booklets within a category (BIB or UBIB) would be administered in equal numbers and without positional bias.

The first purpose was accomplished by forming a cycle of 168 booklets consisting of two sets of BIB booklets (1-57) and nine sets of UBIB booklets (58-63). The BIB and UBIB booklets were merged as follows:

			1	2	58	3	4	59	5	6	60	7	8	61	9	10	62	11
12	63	13	14	58	15	16	59	17	18	19	60	20	21	61	22	23	62	24
25	63	26	27	58	28	29	59	30	31	60	32	33	61	34	35	62	36	37
38	63	39	40	58	41	42	59	43	44	60	45	46	61	47	48	62	49	50
63	51	52	58	53	54	59	55	56	57	60	1	2	61	3	4	62	5	6
63	7	8	58	9	10	59	11	12	60	13	14	61	15	16	62	17	18	19
63	20	21	58	22	23	59	24	25	60	26	27	61	28	29	62	30	31	63
32	33	58	34	35	59	36	37	38	60	39	40	61	41	42	62	43	44	63
45	46	58	47	48	59	49	50	60	51	52	61	53	54	62	55	56	57	63

A given BIB booklet, say #1, appears two times in this cycle; a given UBIB booklet, say #58, appears nine times. Administering this cycle of booklets evenly across the sample of students establishes the ratio of nine UBIB booklets to two BIB booklets.

In a complete cycle of 168 booklets, each of the six UBIB booklets will have appeared nine times and each of the 57 BIB booklets will have appeared two times. As a result of this spiralling, each of the 24 blocks of items used in BIB and UBIB booklets will appear the same number of times in a complete cycle (except for blocks J and R, which are used in both BIB and UBIB booklets at all three grade/age levels, and block Q, which was used in place of block Y for the Grade 4/Age 9 UBIB booklets).

Each block, except for blocks J, R, and Q, appears exactly eighteen times in the 168-booklet cycle. Blocks J and R appear 27 times. Block Q appears 36 times in the Grade 4/Age 9 spiralling cycle. Block Y appears zero times in the Grade 4/Age 9 spiralling cycle.

The second purpose was accomplished by collecting this cycle of 168 booklets into bundles of 23 consecutive booklets, with a subsequent bundle beginning where the previous bundle left off; the last of the 168 booklets was always followed by the first in a continuous circling process (hence the term "spiralling"). As a result, 168 different bundles were created and each booklet distributed evenly throughout 23 positions in the bundles. By shipping consecutive bundles to schools, the likelihood that any given booklet would be used was equalized across the sample.

5.6 The Tape Sample

Four assessment booklets were designed for the tape sample. Each was to be administered to a subsample of 1,250. Each booklet contained the six minute common block and 42 minutes of cognitive exercises and background and attitude items. Since a tape recorder was used in administration, all students in an assessment session were assigned the same booklet.

5.7 Achieved Samples

The results of the implementation of the entire design are shown in Tables 5(6) and 5(7) and Figures 5-1 and 5-2. Table 5(6) presents the number of students assessed by each booklet by grade/age. The same information is graphically depicted in Figure 5-1.

The number of students responding to each BIB and UBIB block appears in Table 5(7) and is graphically depicted in Figure 5-2.

5.8 Advantages and Disadvantages of the Spiral Design

A large, complex assessment design such as that used in the Year 15 NAEP has a number of advantages and disadvantages, which should be mentioned.

5.8.1 Interrelationships Among Exercises

The purpose of the BIB spiral design was to allow the examination of the interrelationships of a large number of exercises, and it does. The final sample includes nineteen 14-minute blocks, 266 minutes in all, of exercises, and for any pair of exercises in these blocks there is a sample of youths who was presented both exercises. Thus, correlations can be computed among all the exercises in this part of the sample. The remaining 112 minutes of exercises are organized so that some, but not all, of the correlations can be calculated.

This design is in contrast to the multiple matrix design which was used previously. Given a fixed sample size, simple matrix sampling and BIB spiralling would administer any particular exercise to the same number of youths, but, by creating more booklets, the spiral design would pair the exercises in a block with many different blocks of exercises, thus increasing the number of comparisons that could be made. Consequently, many correlations are possible, most of them based on a fairly small, though well-selected, sample. In the design as implemented, correlations within a block are based on about 2,000 students for an age or grade separately; correlations between blocks are based on about 222 students for an age or grade separately. In contrast, the simple matrix sampling used

Table 5(6)

Number of Booklets Administered
Spiral and Tape Samples

Booklet Number	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17
1	310	355	346
2	311	346	363
3	316	343	355
4	319	340	354
5	309	346	339
6	320	331	341
7	317	335	335
8	306	342	333
9	315	338	336
10	308	344	342
11	309	343	327
12	309	349	337
13	309	353	324
14	306	344	340
15	309	343	333
16	305	339	337
17	308	336	340
18	296	348	338
19	308	357	340
20	302	343	340
21	312	336	340
22	313	337	347
23	305	328	354
24	307	332	338
25	317	328	336
26	312	325	344
27	315	327	350
28	329	322	347
29	319	328	349
30	317	314	350
31	307	324	338
32	316	332	345
33	306	331	344
34	296	328	344
35	302	335	340
36	311	336	344

Table 5(6)
(continued)

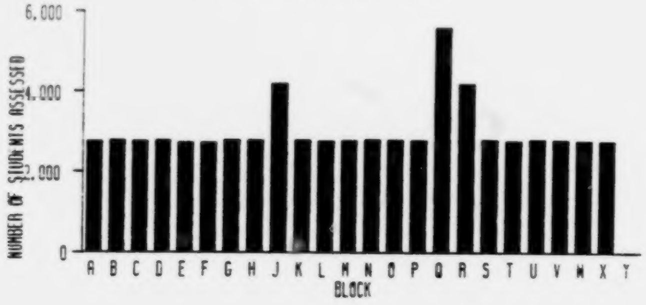
Number of Booklets Administered
Spiral and Tape Samples

Booklet Number	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17
37	303	336	350
38	306	342	345
39	306	346	335
40	291	352	343
41	308	346	348
42	290	345	359
43	292	344	353
44	305	342	349
45	317	352	345
46	313	344	343
47	314	332	338
48	310	330	341
49	317	325	348
50	309	338	351
51	315	323	349
52	322	338	350
53	316	349	356
54	312	350	344
55	316	338	347
56	313	349	329
57	317	339	346
58	1422	1516	1572
59	1396	1527	1537
60	1416	1529	1574
61	1395	1513	1528
62	1385	1520	1528
63	1405	1502	1543
Tape Booklets			
64	1403	1310	1539
65	1356	1276	1540
66	1389	1283	1596
67	1344	1289	1534
Total	31579	33563	35070
Total Spiral	26087	28405	28861
Total Tape	5492	5158	6204

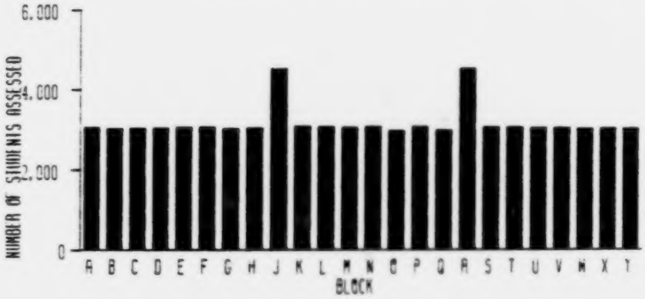
Figure 5-1

BIB Spiral Sample
Number of Students per Block

AGE 9 / GRADE 4 (TOTAL N = 26.087)



AGE 13 / GRADE 8 (TOTAL N = 28.405)



AGE 17 / GRADE 11 (TOTAL N = 28.661)

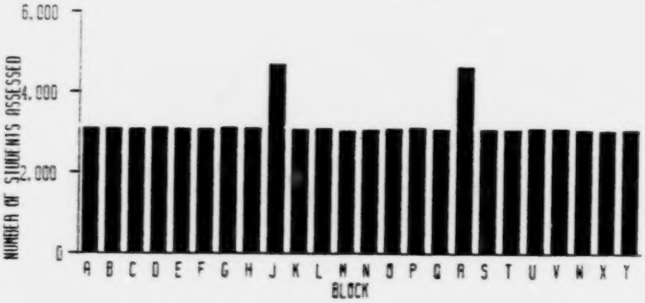


Table 5(7)

Number of Blocks Administered: Spiral and Tape Samples

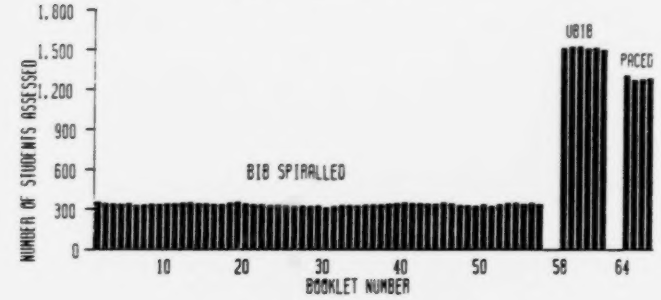
Block	Grade 4/Age 9 Total	Grade 8/Age 13 Total	Grade 11/Age 17 Total
		Spiral Sample	
A	2771	3075	3098
B	2795	3042	3093
C	2776	3052	3084
D	2790	3053	3124
E	2741	3069	3089
F	2744	3072	3082
G	2804	3030	3122
H	2795	3046	3100
J[1]	4213	4525	4701
K	2800	3089	3080
L	2778	3075	3100
M	2792	3057	3045
N	2810	3076	3066
O	2806	2974	3095
P	2803	3075	3120
Q[2]	5611	2982	3083
R[3]	4211	4524	4641
S	2815	3063	3084
T	2788	3060	3076
U	2818	3043	3109
V	2811	3042	3102
W	2790	3022	3071
X	2780	3033	3056
Y[4]		3029	3080
Total Spiral	26087	28405	28861
		Tape Sample	
P64	1403	1310	1539
P65	1356	1276	1540
P66	1389	1283	1596
P67	1344	1289	1534

[1] Block J appeared in both BIB and UBIB
[2] Block Q was substituted for Block Y in books 59 and 63 for Grade 4/Age 9
[3] Block R appeared in both BIB and UBIB booklets
[4] Block Y was not administered at Grade 4/Age 9

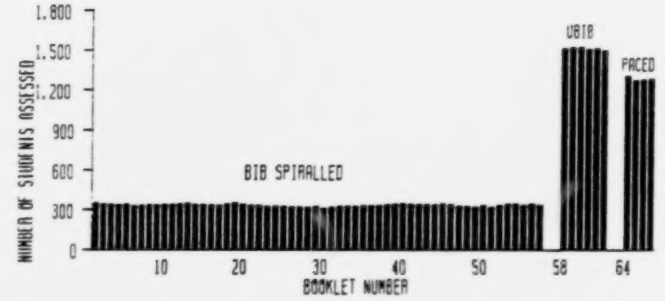
Figure 5-2

BIB Spiral, UBIB Spiral and Tape Samples
Number of Students per Booklet

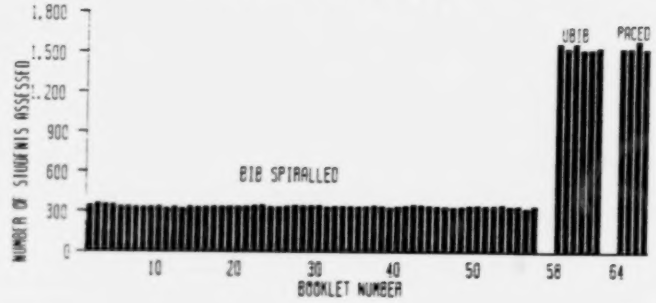
AGE 9 / GRADE 4 (TOTAL N = 31,579)



AGE 13 / GRADE 8 (TOTAL N = 33,563)



AGE 17 / GRADE 11 (TOTAL N = 35,070)



previously allowed the larger number of exercises within a package to be correlated (based on around 2,500 students) but did not allow any calculation of correlations among exercises in different packages.

5.8.2 The Cost of Complexity

Clearly, spiralling is expensive in printing costs as well as in the costs of design talent and detail management. Including the multiple matrix sampling that was done for this NAEP, 67 booklets were created for each of the three grade/ages assessed; thus, there were 201 booklets created in all. It was expensive to produce many booklets in small volumes. It was tedious to manage a task in which every detail had to be multiply checked. Another substantial cost was incurred by the creation of an intelligent data entry system, since developing a way to read the booklets by machine was impossible, given available time and resources.

Spiralling had, however, reduced costs in some ways. The system was robust against failures in the field, since a serious biasing of results by having the exercise administrators use the wrong bundle of booklets was most unlikely and would have very little effect on the design. The absence of the tape recorder reduced costs in both preparation and administration of the assessment. Most importantly, as noted in Section 5.8.4, the spiral design reduced the number of students needed to achieve a fixed standard error, thus allowing us to assess more exercises.

5.8.3 Tape-Recorded Administration

Losing the ability to administer assessments by tape recorder was not something that the NAEP staff wanted, but came about because of the spiral design. It is clear that, when each student in an assessment session is taking a different booklet, the administration cannot be presented with a single tape recorder. We did not consider tape recorders with headphones for use by individual students.

The advantage of tape-recorded administration is that it allows the separation of reading ability from the subject area being assessed. In a reading assessment, the instructions are tape recorded and the progress through the assessment is paced, although the reading exercises themselves are, of course, not read. In other subject areas, the exercises are read aloud so that students can respond to an exercise even though they may not be able to read it. This is clearly a desirable feature. Additionally, the pacing feature of tape-recorded administration tends to ensure that each student is exposed to each exercise. This is also a desirable feature.

And yet, the utility of the NAEP is greatly enhanced by developing exercises that teachers or local or state personnel can readily administer to their students, the results of which can then be compared to the NAEP sample. Teachers are not likely to simulate the tape recording; thus, any comparisons would be suspect. We know of no local or state assessments that currently use tape-recorded administrations (although some states and districts in the early days of NAEP replicated all administration procedures including taped administrations). Thus, the tape recorder had the effect of setting the NAEP results apart from all other student assessments.

5.8.4 Sampling Efficiency

One advantage of the spiral design is that it presents a particular block of exercises to fewer persons in a school, but to more schools. In this way, the cluster effect is markedly reduced; thus, the students are used more efficiently. Given reasonable assumptions, it has been estimated that the required sample size to achieve a given standard error is reduced by about 20 to 25 percent by BIB spiralling, as compared to multiple matrix sampling; alternatively, the standard errors could be reduced by about 10 to 15 percent if the sample size were kept constant (Hansen, Tepping, Lago, & Burke, 1984). Analyses of the design effects from the Year 15 NAEP (discussed in Chapter 14.2) show that this reduction in variability has, indeed, taken place.

5.8.5 Statistical Issues

As mentioned above, spiralling does not result in a complete, rectangular data matrix that can be analyzed using standard statistical systems nor does it generate data which are consistent with normal statistical methods. The techniques used to analyze such a dataset are discussed in subsequent chapters.

The exercise assignment procedures produced a total of 67 different samples of the population of students of a particular grade/age, one for each of the 63 BIB/UBIB spiral booklets and one for each of the four tape booklets. Although each of these samples involved different students, they are, in a particular sense, equivalent to each other. Because they were selected by probability sampling techniques (described in Chapter 4), the complete set of students of a given grade/age who were selected for assessment are a representative probability sample of the population of students of that grade/age designation. The procedure for designating whether a given student was to be assessed in a spiral session or in a tape session and, if a tape session, which of the four booklets was to be used, was also done in a (controlled) random manner; the procedure (given in Chapter 4) ensured that every student could have been selected for any one of the four tape sessions or for a spiral session. This random assignment was controlled (by systematic selection) to ensure that each of the five samples (the four tape samples and the combined spiral sample) was representative of the population, in particular controlling for all of the stratification variables (region, size and description of community) as well as the size of the school.

The (larger) spiral sample was further divided into 63 subsamples by the BIB/UBIB spiral technique described previously. As in the case of the assignment of type of session to student, this division was also done in a systematic (but random) manner, to ensure that every student who was selected for a spiral session could have received any one of the 63 booklets. This random assignment was done within sessions within each school and so is more likely to result in samples of students which closely match each other in terms of their demographic characteristics.

The 63 samples corresponding to the spiral booklets and the four samples corresponding to the tape booklets given at a particular grade/age are each representative samples of their target population of all students in the grade/age. Since any assessed student could have been placed in any one of these samples, and because of the balance that is enforced by the method of sampling, each of these samples can be deemed equivalent, in a sense, to each other. We will call them randomly equivalent. Because of the closer match between the various samples that is possible with spiralling, the equivalence between the spiral samples is closer than is the equivalence between the tape samples.

Chapter 6

INSTRUMENT AND ITEM INFORMATION[1]

Janet R. Johnson

Educational Testing Service

The Year 15 assessment incorporated four distinct types of instruments: student assessment booklets, a questionnaire for excluded students, a teacher questionnaire, and a school characteristics and policy questionnaire. The data collected from these instruments are available on the public-use data tapes. This chapter begins with a discussion of how cognitive and non-cognitive items were organized into blocks to create the student assessment booklets. Sections 6.1.2 through 6.1.4 provide an overview of the items. The last three sections describe the questionnaires.

6.1 Student Assessment Instruments

Student assessment booklets were composed of items that were either cognitive or non-cognitive. Cognitive items were reading exercises, study skill exercises or writing exercises. Non-cognitive items asked questions relative to the background and attitudes of students. Some non-cognitive items were presented to every student and were placed together in a block called the common block or common core. Others were placed at the beginning of the blocks containing the cognitive items. Later sections of this chapter provide greater detail about both the cognitive and non-cognitive items.

Based upon the Balanced Incomplete Block (BIB) and Unbalanced Incomplete Block (UBIB) sampling design (described in Chapter 5), cognitive and non-cognitive items were grouped into blocks. Twenty-four blocks of items were used to create a total of 63 spiral assessment booklets and four tape-administered booklets for each grade/age.[2] Tables 6(1) and 6(2) show the blocks contained in each booklet used for each grade/age.

---

[1] Some of the tables for this chapter were generated by David Freund and Alfred Rogers; details regarding block assemblage were provided by Kalle Gerritz; and the taxonomy provided in Table A(2) of Appendix A was created by Gita Wilder.

[2] Tape-administered booklets were used in group administrations to "pace" students through booklets with audio recordings. The instructions were read by an announcer; reading passages, items, and response choices

Table 6(1)

Booklet Contents by Block for Grade 4/Age 9

Booklet	Block	Booklet	Block	Booklet	Block
1	TGL	26	TFQ	51	GRO
2	ALP	27	CKG	52	SGD
3	DAT	28	OJS	53	HPN
4	CSE	29	QOD	54	TEK
5	CAH	30	BQJ	55	MGN
6	GFH	31	OTH	56	ANQ
7	KRN	32	BML	57	GJP
8	RMF	33	CRO	58	U*J
9	ONL	34	GOE	59	U*Q
10	FDB	35	SQM	60	V*R
11	EMA	36	BAO	61	V*X
12	SHB	37	KGA	62	W*X
13	MKD	38	OFK	63	W*Q
14	TNJ	39	PST	64	tape
15	MTC	40	EFL	65	tape
16	CLQ	41	HMJ	66	tape
17	HER	42	JED	67	tape
18	CPF	43	FJA		
19	LSK	44	BGC		
20	NBE	45	PBK		
21	NCD	46	SFN		
22	QKH	47	PQE		
23	LHD	48	BRT		
24	ASR	49	PMO		
25	LJR	50	RPD		

*double-length block

Table 6(2)

Booklet Contents by Block for Grade 8/Age 13 and Grade 11/Age 17

Booklet	Block	Booklet	Block	Booklet	Block
1	TGL	26	TFQ	51	GRQ
2	ALP	27	CKJ	52	SGD
3	DAT	28	OJS	53	HPN
4	CSE	29	QOD	54	TEK
5	CAH	30	BQJ	55	MGN
6	GFH	31	OTH	56	ANQ
7	KRN	32	BML	57	GJP
8	RMF	33	CRO	58	U*J
9	ONL	34	GOE	59	U*Y
10	FDB	35	SQM	60	V*R
11	EMA	36	BAO	61	V*X
12	SHB	37	KGA	62	W*X
13	MKD	38	OFK	63	W*Y
14	TNJ	39	PST	64	tape
15	MTC	40	EFL	65	tape
16	CLQ	41	HMJ	66	tape
17	HER	42	JED	67	tape
18	CPF	43	FJA		
19	LSK	44	BGC		
20	NBE	45	PBK		
21	NCD	46	SFN		
22	QKH	47	PQE		
23	LHD	48	BRT		
24	ASR	49	PMO		
25	LJR	50	RPD		

*double-length block

For Grade 4/Age 9, 20 single- and 3 double-length blocks were used to create booklets. For Grade 8/Age 13 and Grade 11/Age 17, 21 single- and 3 double-length blocks were used to create booklets. Tables 6(3), 6(4), and 6(5) show the contents of the blocks and the booklets in which they were placed. Each single-length block contained fourteen minutes of assessment items. Approximately, the first two minutes were devoted to background and attitude items while the remainder of the fourteen minutes contained cognitive items. The double-length blocks were similarly arranged but allowed 28 minutes total assessment time, the majority of which was to be devoted by the student to responding to the cognitive items. It is important to remember that while the content of some blocks was identical for more than one grade/age, and sometimes identical for all three grade/ages, this was not true in every instance. For example, the cognitive items contained in Block X for Grade 4/Age 9 are entirely different from those contained in Block X for Grade 8/Age 13 and Grade 11/Age 17. As illustrated by the tables, and described below, different blocks contained different types of items.

Blocks A through G were writing blocks, which contained writing-related non-cognitive items followed by writing exercises. Blocks H through R were reading blocks, which contained both general and reading-related non-cognitive items followed by reading exercises. The number of reading or writing exercises within a block, listed in Tables 6(3) through 6(5) under the heading "Cog. Items", varied from one block to another.

Some items that had been considered reading items in the broader definition used by the learning area committees of earlier assessments were re-classified as "study skill" items for Year 15. An item was classified as a study skill item if it required some specially learned skill above and beyond the facility of recognizing and understanding the printed word. For example, these items included those whose stimulus was a bar graph, a telephone bill or a table of contents. Study skill items were concentrated in blocks S and T; some study skill items also appeared in the four tape booklets. They were excluded from the group of items used in the IRT analysis (see Chapter 10.3) because they were believed to be representative of a different dimension.

Blocks U, V, and W contained a combination of writing and reading items and were 28 minutes long. Block X was fourteen minutes long and contained both reading and writing items. Block Y, which was not administered to Grade 4/Age 9, was fourteen minutes long and contained reading items. Blocks X and Y were used exclusively in combination with the 28-minute blocks.

Tables 6(3) through 6(5) also provide the total number of each type of item--background, writing or reading--for each age. As can be seen, the item pool varied in number of items from one grade/age to another.

---

were read by the student. The taped administrations were used in previous NAEP assessments and were used again in Year 15 to explore the effects of the change from audiotaped recordings to pencil-and-paper instruments.

Table 6(3)

Assessment Items for Grade 4/Age 9

Block	Type	Bg. Items	Writing Items	Reading Items	No. Total Items	No. Cog. Items	Booklets Containing Block
Common	Bg.	1-37			37		1-63
A	Wr.	1-12	13		13	1	2 3 5 11 24 36 37 43 56
B	Wr.	1-15	16		16	1	10 12 20 30 32 36 44 45 48
C	Wr.	1-22	23		23	1	4 5 15 16 18 21 27 33 44
D	Wr.	1-24	25		25	1	3 10 13 21 23 29 42 50 52
E	Wr.	1-9	10,11		11	2	4 11 17 20 34 40 42 47 54
F	Wr.	1-5	6,7		7	2	6 8 10 18 26 38 40 43 46
G	Wr.	1-6	7,8		8	2	1 6 34 37 44 51 52 55 57
H	Rdg.	1-4		5-15	15	11	5 6 12 17 22 23 31 41 53
J	Rdg.	1-11		12-24	24	13	14 25 27 28 30 41 42 43 57 58
K	Rdg.	1-8		9-19	19	11	7 13 19 22 27 37 38 45 54
L	Rdg.	1-19		20-26	26	7	1 2 9 16 19 23 25 32 40
M	Rdg.	1-4		5-16	16	12	8 11 13 15 32 35 41 49 55
N	Rdg.	1-11		12-25	25	14	7 9 14 20 21 46 53 55 56
O	Rdg.	1-11		12-22	22	11	9 28 29 31 33 34 36 38 49
P	Rdg.	1-6		7-19	19	13	2 18 39 45 47 49 50 53 57
Q	Rdg.	1-9		10-21	21	12	16 22 26 29 30 35 47 51 56 59 63
R	Rdg.	1-4		5-16	16	12	7 8 17 24 25 33 48 50 51 60
S	St.Sk.	1-18		19-33	33	15	4 12 19 24 28 35 39 46 52
T	St.Sk.	1-18		19-35	35	17	1 3 14 15 26 31 39 48 54
U	Comb.	1-17	18	19-27	27	10	58 59
V*	Comb.	1-28	36	29-35	36	8	60 61
W	Comb.	1-36	39,43	37-38,40-42	43	7	62 63
X	Comb.	1-15	16	17-20	20	5	61 62
Total Cognitive			15	173		188	

*Item 35 is a three-part reading item

Table 6(4)

Assessment Items for Grade 8/Age 13

Block	Type	Bg. Items	Writing Items	Reading Items	No. Total Items	No. Cog. Items	Booklets Containing Block
Common	Bg.	1-37			37		1-63
A	Wr.	1-12	13		13	1	2 3 5 11 24 36 37 43 56
B	Wr.	1-15	16		16	1	10 12 20 30 32 36 44 45 48
C	Wr.	1-22	23		23	1	4 5 15 16 18 21 27 33 44
D	Wr.	1-24	25		25	1	3 10 13 21 23 29 42 50 52
E	Wr.	1-9	10,11		11	2	4 11 17 20 34 40 42 47 54
F	Wr.	1-5	6,7		7	2	6 8 10 18 26 38 40 43 46
G	Wr.	1-6	7,8		8	2	1 6 34 37 44 51 52 55 57
H	Rdg.	1-5		6-18	18	13	5 6 12 17 22 23 31 41 53
J	Rdg.	1-10		11-24	24	14	14 25 27 28 30 41 42 43 57 58
K	Rdg.	1-8		9-17	17	9	7 13 19 22 27 37 38 45 54
L	Rdg.	1-21		22-27	27	6	1 2 9 16 19 23 25 32 40
M	Rdg.	1-4		5-16	16	12	8 11 13 15 32 35 41 49 55
N	Rdg.	1-11		12-23	23	12	7 9 14 20 21 46 53 55 56
O*	Rdg.	1-11		12-21	21	10	9 28 29 31 33 34 36 38 49
P	Rdg.	1-6		7-15	15	9	2 18 39 45 47 49 50 53 57
Q	Rdg.	1-6		7-23	23	17	16 22 26 29 30 35 47 51 56
R	Rdg.	1-4		5-19	19	15	7 8 17 24 25 33 48 50 51 60
S	St.Sk.	1-18		19-37	37	19	4 12 19 24 28 35 39 46 52
T	St.Sk.	1-18		19-38	38	20	1 3 14 15 26 31 39 48 54
U	Comb.	1-17	18	19-31	31	14	58 59
V	Comb.	1-28	32	29-31	32	4	60 61
W	Comb.	1-36	37,42	38-41	42	6	62 63
X	Comb.	1-15	16	17-24	24	9	61 62
Y	Comb.	1-3		4-10	10	7	59 63
Total Cognitive			15	191		206	

*Item 15 is a two-part reading item

Table 6(5)

Assessment Items for Grade 11/Age 17

Block	Type	Bg. Items	Writing Items	Reading Items	No. Total Items	No. Cog. Items	Booklets Containing Block									
Common	Bg.	1-48			48		1-63									
A	Wr.	1-12	13		13	1	2	3	5	11	24	36	37	43	56	
B	Wr.	1-15	16		16	1	10	12	20	30	32	36	44	45	48	
C	Wr.	1-22	23		23	1	4	5	15	16	18	21	27	33	44	
D	Wr.	1-24	25		25	1	3	10	13	21	23	29	42	50	52	
E	Wr.	1-9	10,11		11	2	4	11	17	20	34	40	42	47	54	
F	Wr.	1-5	6,7		7	2	6	8	10	18	26	38	40	43	46	
G	Wr.	1-6	7,8		8	2	1	6	34	37	44	51	52	55	57	
H	Rdg.	1-6		7-19	19	13	5	6	12	17	22	23	31	41	53	
J	Rdg.	1-11		12-17	17	6	14	25	27	28	30	41	42	43	57	58
K	Rdg.	1-8		9-17	17	9	7	13	19	22	27	37	38	45	54	
L	Rdg.	1-26		27-32	32	6	1	2	9	16	19	23	25	32	40	
M	Rdg.	1-4		5-16	16	12	8	11	13	15	32	35	41	49	55	
N	Rdg.	1-20		21-32	32	12	7	9	14	20	21	46	53	55	56	
O	Rdg.	1-11		12-24	24	13	9	28	29	31	33	34	36	38	49	
P	Rdg.	1-14		15-25	25	11	2	18	39	45	47	49	50	53	57	
Q	Rdg.	1-6		7-17	17	11	16	22	26	29	30	35	47	51	56	
R	Rdg.	1-11		12-20	20	9	7	8	17	24	25	33	48	50	51	60
S	St.Sk.	1-18		19-37	37	19	4	12	19	24	28	35	39	46	52	
T	St.Sk.	1-18		19-38	38	20	1	3	14	15	26	31	39	48	54	
U	Comb.	1-17	18	19-31	31	14	58	59								
V	Comb.	1-37	41	38-40	41	4	60	61								
W	Comb.	1-38	39,44	40-43	44	6	62	63								
X	Comb.	1-15	16	17-24	24	9	61	62								
Y	Comb.	1-5		6-12	12	7	59	63								
Total Cognitive			15	176		191										

6.1.1 Assembling Reading and Writing Items into Blocks

The following considerations were taken into account during the process of assembling the blocks:

(1) Because of the order of assessment administration, blocks for Grade 8/Age 13 were developed first, then those for Grade 4/Age 9, and finally those for Grade 11/Age 17. Ideally, blocks for all three grade/ages should have been developed together.

(2) An item was selected to be placed within a specific block based on the time required to complete the item.

(3) For Grade 11/Age 17, some blocks were repeated intact from the blocks assembled for Grade 8/Age 13.

(4) In general, an attempt was made to start blocks with easy items and progress to difficult ones. This was not always possible.

(5) When a reading item required a lengthy written response, the item was always placed at the end of a block.

(6) Whenever possible, reading items were physically arranged so that the reading passage and the items appeared on the same or facing pages. This was not possible when the stimulus material was lengthy.

(7) Any item that had been revised and was, therefore, different from its earlier form as used in previous assessments was considered to be a new item.

(8) The tapes contained only items that had been used in past assessments. Items were fit into the tape blocks based on the timing of the items as taken from the tape scripts.

6.1.2 Reading Items

The reading items included short and long reading passages, graphically presented materials, poems, and reference materials (e.g., tables of contents). Some items required a multiple-choice response, some open-ended items required a brief written response, and some required written essays. These latter items, of which there was a total of twelve across the three grade/ages, were professionally scored. (The professional scoring process is discussed in Chapter 8.2.)

Some of these items had been developed for the earliest reading assessment and re-used in some or all of the subsequent assessments; some items had been developed exclusively for the 1983-84 assessment; and some items had been developed and used only once over the years. (Development of the reading objectives and items is discussed in Chapter 3.) In addition, some items had remained unchanged in wording and arrangement while others had undergone a variety of alterations. Each item was carefully researched as to its use and possible alteration over time. This process became important specifically for those items included in the tape booklets for use in the Year 15 trend analysis. (See Table B(1) in Appendix B for a list of the items initially considered for use in the trend analysis.)

Tables 6(6) through 6(8) examine the tapes for each age. These tables show which items (by item location number) from each spiral block were used in the assembly of the tape booklets. For Age 9, Tape 2 contains one reading item that does not appear within any of the spiral blocks. For Age 17, there are 20 such reading items across all four tapes.

Table A(1) in Appendix A provides a complete descriptive list of all Year 15 reading items with their corresponding block or tape numbers and item numbers. As can be seen from this table, the number of items presented to each age varied. Some items overlapped all three grade/ages, some overlapped two grade/ages, and some were particular to a grade/age. A total of 176 reading items was presented to Grade 4/Age 9; a total of 192 reading items was presented to Grade 8/Age 13; and a total of 196 reading items was presented to Grade 11/Age 17. Complete item text is available on the microfiche that accompanies the public-use data tapes.

6.1.3 Writing Items

Writing items appeared in spiral blocks A through G and U through X and one or two of the tapes depending upon the age group. Table 6(9) presents the pool of writing items and their block or tape locations by age.

From a total pool of 22 writing items, 15 were used for each grade/age. Some of these items had been used in one or both of the previous writing assessments. By design, students who received one or more writing blocks could be asked to respond to as few as one writing item or as many as four.

The writing items were developed to assess performance in three writing areas: informative, persuasive and imaginative. Students were asked to write, for example, letters, descriptive essays, or narrative pieces. (The development of writing objectives and items is discussed in Chapter 3; professional scoring of the writing responses is discussed in Chapter 8.2.)

Four of the writing items, Dali, Aunt May, Split Session and Hole in the Box, appeared in the tape booklets to be used in determining writing trends. For more information concerning writing trends see Writing: Trends Across the Decade, 1974-84 (Applebee, Langer, & Mullis, 1986a).

Table 6(6)
Cognitive Items from Spiral Blocks on Age 9 Tapes

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S+	T+	U	V	W	X	Y	Total Items on Tape
Tape 1											21		17 18 19 20 21 22			10 11 12	9	19 20 26 27 28 29	19 20 21 22 23 26 28 29 30 31 32 33 34 35	24 25 26 27	30				36
Tape 2	13*							6 7 8		9 10 11	22	13		15	10 11 12 13	15	10						16*		18 One item on tape does not appear within the blocks
Tape 3								5 10 11 12 13 14 15	12 13 14 15 16 17	16 17		10 11 12		16		13 14 16	11 12	21 22 23 24 25							29
Tape 4									19 20	18 19				12 13 14 17 18 19	7 8 9			30		18*	29	37 38			18

+ Study Skills Items
* Writing Items

Table 6(7)
Cognitive Items from Spiral Blocks on Age 13 Tapes

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S†	T†	U	V	W	X	Y	Total Items on Tape
Tape 1													12			7		22	19	19					34
													13			8		30	20	20					
													14			9		31	21	21					
													15					32	22	22					
													19					33	23	23					
													20					34	26	24					
																		35	28						
																		36	29						
																		37	30						
																			31						
Tape 2	13*								17	9	23		18			12							16*		17
									18	10						13									
									19	11						14									
										14						15									
										15															
										16															
Tape 3								7	11			11	16	16		10		19	32					4	30
								8	12			12				11		20	33					5	
								12	13			13						21	34					6	
									14										35					7	
									15															8	
									16															9	
																								10	
Tape 4									20	12				12	7	16			27	18*			17		26
									21	13				13	8	17							18		
														14	9	18							19		
														17									20		
														18									21		
														19									22		
																							23		
																							24		

† Study Skills Items
* Writing Items

Table 6(8)
Cognitive Items from Spiral Blocks on Age 17 Tapes

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S†	T†	U	V	W	X	Y	Total Items on Tape
Tape 1													21					22	19	19					34
													22					30	20	20					Three items on tape do not appear within the blocks
													23					21	21	21					
													24					32	22	22					
													28					33	23	23					
													29					34	26	24					
																		35	28						
																		36	29						
																		37	30						
																			31						
Tape 2	13*								12	9	27		27			7							16*		17
									13	10						8									One item on tape does not appear within the blocks
									14	11						9									
										14															
										15															
										16															
Tape 3								13				11	25			10		19	32					6	30
												12						20	33					7	Ten items on tape do not appear within the blocks
												13						21	34					8	
																			35					9	
																								10	
																								11	
																								12	
Tape 4										12				12		11			27	18*			17		26
										13				13		12							18		Six items on tape do not appear within the blocks
														14		13							19		
														21									20		
														22									21		
																							22		
																							23		
																							24		

† Study Skills Items
* Writing Items

Table 6(9)

Year 15 Writing Items

		Block and Tape Locations		
	Item	Age 9	Age 13	Age 17
N000102	DALI	A,Tape 2	A,Tape 2	A,Tape 2
N000202	SCHOOL RULE	B	B	B
N000302	RECREATION OPP.		C	C
N000402	FOOD ON FRONTIER	D	D	D
N000502	DISSECTING FROGS		E	
N000602	XYZ COMPANY	E	E	
N000702	SWIMMING POOL	F	F	F
N000802	PETS	F	F	
N000902	RADIO STATION	G	G	
N001002	APPLEBY HOUSE	G	G	G
N007202	HOLE IN THE BOX	U,Tape 4	U,Tape 4	U,Tape 4
N007602	FLASHLIGHT	V	V	V
N007702	GHOST STORY	W	W	W
N007902	FAVORITE MUSIC	W	W	W
N008002	SPLIT SESSION		X,Tape 2	X,Tape 2
N014702	PLANTS	C		
N014802	SPACESHIP	E		
N014902	AUNT MAY	X,Tape 2		
N018002	SPACE PROGRM			E
N019002	JOB APPLICATION			E
N020002	UNCLE			F
N021002	BIKE LANE			G

Tables 6(6) through 6(8) examine the tapes for each age. These tables show which items (by item location number) from each spiral block were used in the assembly of the tape booklets. Writing items are indicated by asterisks.

Complete writing item text is available on the microfiche that accompanies the public-use data tapes.

6.1.4 Non-Cognitive Items

For the Year 15 assessment, each spiral and tape booklet included six minutes of background and attitude items common to all students. These items are general questions concerning materials in the home, parental education, etc. Additional background and attitude items were spiralled throughout BIB and UBIB booklets (see Chapter 5). These attitude items related to objectives formulated for reading and writing. The items measured students' perceptions of their teachers' instructional practices in reading and writing; their own study habits and reading activities; their perceptions of the value of reading and writing; and their assessment of themselves as readers and writers.

Table A(2) in Appendix A lists descriptors of all of the background and attitude items grouped by the topics they were designed to address. The series of a letter and numbers preceding each descriptor is the unique NAEP ID assigned to that particular item. If the ID begins with "B", the item was included in the common block of items administered to all students. If the ID begins with "S", the item appeared at the beginning of a single- or double-length block.

Table A(3) in Appendix A lists descriptors of all background and attitude items in NAEP ID order with block location and item number within block for each grade/age. The common block (CB) items are listed first, followed by items which appear in the single- and double-length blocks (A through X). Grade 11/Age 17 students were presented an additional number of items, many of which were curricula-specific. Complete item text is available on the microfiche that accompanies the public-use data tapes.

In addition to the common core items, the Year 15 tape booklets contained additional non-cognitive items, which appeared as a separate section at the end of the booklets. These items were drawn from the pool of items appearing in the spiral booklets.

6.2 The Excluded Student Questionnaire

The Excluded Student Questionnaire was developed and used for the first time in the Year 15 assessment. It was designed to gather more information about particular conditions for exclusion and characteristics of the learning experience of excluded students.

The questionnaire was completed by school personnel for every student who was selected for inclusion in the NAEP sample but was unable to respond to items because he or she was judged by school personnel to be non-English speaking, educable mentally retarded or functionally disabled. The four-page questionnaire was used to gather information concerning special education, language, and other student programs. A copy of the Excluded Student Questionnaire is available on the microfiche that accompanies the public-use data tapes.

Of the 104,437 students sampled for the Year 15 assessment, 4,225 were ineligible or excluded by the school due to classification as educable mentally retarded, non-English speaking, or functionally disabled. There were 1,416 (4.3 percent) excluded students in Grade 4/Age 9, 1,448 (4.1 percent) in Grade 8/Age 13 and 1,361 (3.7 percent) in Grade 11/Age 17.

6.3 The Teacher Questionnaire

The Teacher Questionnaire was developed and used for the first time in Year 15. It was designed to gather information on the curricula and teaching methods used by selected English and Language Arts teachers. The data were provided by teachers who completed a nine-page questionnaire which included questions concerning years of teaching experience, frequency of writing assignments, teaching materials used, the availability and use of computers, and perceptions of the school and its curricula.

To sample teachers for the teacher questionnaire we associated the student's main language arts or English teacher with each student participating in the spiral assessment. We requested that the student's main English teacher be identified in the background information sheet. The sample of teachers was then drawn by selecting one student from each of the sessions being conducted at the school. Each teacher sampled received only one questionnaire even though he or she may have been associated with more than one of the students subsampled for this purpose. Further detail on the sampling of teachers is provided in Section 2.4 of Chapter 2.

Responses were received from a total of 1,027 fourth grade teachers, 790 eighth grade teachers and 915 eleventh grade teachers.

Three versions of the Teacher Questionnaire were developed--one for each grade/age. A copy of each Teacher Questionnaire is available on the microfiche that accompanies the public-use data tapes.

6.4 The School Characteristics and Policy Questionnaire

A School Characteristics and Policy Questionnaire was distributed to each participating school to be completed by either the school's principal or another person familiar with data concerning enrollment, facilities, curricula and staff development.

The five-page questionnaire was developed for two purposes: to collect school data proven by research studies to be related to student performance; and to collect school data for use by educational policymakers both to monitor implementation of existing policies and to identify new policy issues.

The questionnaire items were grouped according to eight categories: principal, students, staff, standards, program, computers, school climate, and school finance.

Responses were received from 663 fourth-grade schools, 486 eighth-grade schools, and 331 eleventh-grade schools. Cooperation rates were 88.6 percent, 90.3 percent, and 83.9 percent for fourth-grade, eighth-grade, and eleventh-grade schools, respectively; the overall cooperation rate was 88.1 percent.

Because no eligible students were selected in several schools that submitted responses, the number of schools for which data are retained in the NAEP database is less than the number of schools from which responses were received. The NAEP database contains data for 661 fourth-grade schools, 478 eighth-grade schools, and 326 eleventh-grade schools.

A copy of the School Characteristics and Policy Questionnaire is available on the microfiche that accompanies the public-use data tapes.

Chapter 7

FIELD ADMINISTRATION

Renee Slobasky
Nancy Caldwell

Westat, Inc.

As a subcontractor to ETS, Westat, Inc. was responsible for field activities leading to and including administration of the assessment sessions and delivery of completed assessment booklets to ETS. (Westat, Inc. was also responsible for sample design and implementation, discussed in Chapter 4.) This chapter describes the Westat field organization and operations for the Year 15 assessment. Details of field administration activities are available in the Westat Report on Field Operations and Data Collection Activities - NAEP Year 15 (1984).

The Year 15 assessment focused on the learning areas of reading and writing. For this assessment, over 1,600 schools were sampled and invited to cooperate. Of this number, 1,465 schools actually participated. Within these schools, a sample of 114,075 students was selected to be assessed.

7.1 Schedule of Year 15 Field Activities

The Year 15 pre-assessment and assessment field activities were conducted from May 1983 to May 1984. The period from May to September 1983 was devoted to the pre-assessment activities of establishing the field force and developing all materials and procedures to be used during the assessments. Pre-assessment activities are described in Section 7.2.

In early October 1983, the assessment began. Thirteen-year-olds and eighth graders were assessed during the period from October 10 to December 16, 1983. Nine-year-olds and fourth graders were assessed from January 2 to March 9, 1984. The last group, the seventeen-year-olds and eleventh graders, were the focus of assessment activities from March 12 to May 11, 1984. (In four schools, makeup sessions were scheduled after May 11 due to poor attendance at the initial sessions.) Conduct of the assessments is described in Section 7.3.

Quality control was an important part of the entire field effort. In addition to the field monitoring activities described in Section 7.4, in-person site visits were made to a sample of schools and an additional sample of schools was interviewed by telephone. These quality control activities are described in Section 7.3.5.

7.2 Pre-Assessment Activities

7.2.1 Establish Field Organization

The home office staff involved in supervising the field operations included the field director and assistant field director. The field director coordinated all field operation activities in the home office and had the primary reporting relationship with half of the district supervisors. The assistant field director had the primary reporting relationship with the other half of the district supervisors, and was also responsible for materials distribution and directing the receipt of reporting forms (as described in Section 7.4) in the home office.

As described in Chapter 4, Sample Selection and Instrument Collection, 64 counties or groups of counties (primary sampling units, or PSUs) were selected for the Year 15 sample. The 64 PSUs were then grouped into sixteen major regions based on a fairly equal geographic spread of schools. In June 1983, Westat recruited sixteen district supervisors and five alternate supervisors to assist district supervisors when there were scheduling conflicts. Each district supervisor was assigned one of the sixteen major regions.

The district supervisor was responsible for a variety of tasks. During the pre-assessment phase, the district supervisor contacted school districts (after the initial contact was made by ETS) and arranged for an introductory meeting with school personnel; conducted the introductory meeting and scheduled each school's assessment; and recruited exercise administrators to assist in the conduct of the assessments.

During the assessment phase, the district supervisor sampled the students to be assessed in each school; trained and provided support to the exercise administrator who conducted the assessment; distributed and collected the Excluded Student Questionnaires, Teacher Questionnaires, and School Characteristics and Policy Questionnaire; and completed all administrative reporting forms. After an assessment at a school was complete, the district supervisor packed and shipped all school materials to ETS.

Each district supervisor hired between one and three exercise administrators per PSU. A few exceptions were made in regions where the schools in several PSUs were clustered in large metropolitan areas. In these regions, the supervisors hired three to four exercise administrators who worked the entire metropolitan area. The exercise administrators assisted the supervisor in selecting the sample of students to be assessed, conducted the assessment sessions, and prepared completed exercise booklets for shipping.

For the most part, staffing of the field organization remained fairly constant throughout the field period. One district supervisor was replaced prior to the assessment phase of the project. There was approximately

15 percent attrition among the exercise administrators; however, this turnover had little, if any, impact on the conduct of the work.

The background and experience of the district supervisors are summarized in Table 7(1). As can be seen from the figure, fourteen supervisors lived in one of the PSUs of their region, four had worked on NAEP before (two as supervisors, two as exercise administrators), all had had some supervisory experience, ten had worked for Westat before the NAEP project, and eight had an educational background (teaching or educational research).

7.2.2 District Supervisor Training

District supervisors' training was divided into two parts. Part I, which lasted two days, introduced the study and explained pre-assessment activities. Part II, which lasted three days, was devoted to actual assessment activities. Conducting the training in two short sessions rather than one long one was a departure from past practice. With two sessions, each session could focus on the particular tasks at hand and not present too much detailed information at once. This arrangement also gave Westat home office staff more time to assess the strengths and weaknesses of the supervisors and to take any necessary corrective action.

The first supervisors' training session was held August 1-2, 1983. Training was conducted by the Westat project director and field director, with introductory remarks and explanatory notes made by ETS. In attendance were the supervisors, alternate supervisors, and representatives from ETS' regional offices who were to make the initial contacts with school districts to solicit participation. The topics included an overview of NAEP and the supervisors' responsibilities; procedures for contacting schools and conducting introductory meetings; planning the schedule of assessments within PSUs; and recruiting and training exercise administrators.

Immediately prior to the assessment phase of the field effort, Westat and ETS staffs reassembled for Part II of supervisor training, held October 3-5, 1983. Topics included training and supervising exercise administrators; the student sample selection process; administrative procedures for conducting the assessments; supervisory responsibility for quality control of assessment sessions and all NAEP materials; and procedures for shipping materials and reports to Westat and ETS. This session was also attended by Westat staff and the ETS regional staff who would be conducting in-person quality control visits to sampled schools to verify the sampling and observe the supervisors and exercise administrators at work.

Table 7(1)

Criteria Met by NAEP Supervisors by Supervisory Region

Region	Lived Within Selected PSU	Prior NAEP Experience	Prior Westat Experience	Prior Supervisory Experience with Westat or Other Research Organization	Prior Employment in Education
I			X	X	X
II	X	X	X	X	
III	X		X	X	
IV	X		X	X	
V	X	X		X	X
VI	X			X	X
VII	X		X	X	
VIII	X	X		X	
IX	X			X	X
X	X	X	X	X	
XI	X		X	X	X
XII	X			X	
XIII	X		X	X	
XIV			X	X	X
XV	X		X	X	X
XVI	X			X	X

7.2.3 Solicit Cooperation of School Districts and Sample Schools

7.2.3.1 Preliminary Contacts

During June, July, and August 1983, ETS and Westat notified the appropriate state and local school officials about NAEP and requested the cooperation of the sample schools. The activities during these three months are discussed in detail below.

Recruiting of schools for NAEP actually began in June, once the sample of schools had been selected and their corresponding school districts identified. ETS contacted the chief state school officers in each state and requested them to notify the school district superintendents. In July, ETS sent a letter to the superintendents and heads of private schools inviting their participation. Under separate cover, informational material on NAEP and, if applicable, a list of the original sample schools in the district, were also sent. These initial contacts, which were completed prior to supervisor training, paved the way for the telephone contacts to follow.

Immediately after training, ETS regional staff contacted the superintendents to discuss NAEP further and to obtain their cooperation. The results of these contacts were documented on the Results of Contact Form. Once cooperation had been determined, ETS staff mailed two copies of this form to the district supervisor and one copy to the Westat home office.

Upon receipt of the Results of Contact Form, the district supervisor called the contact person listed on the form to arrange for an introductory meeting with representatives of the sample schools and to obtain updated information on schools in the district. Any new school openings, school closings or changes in grade or enrollment were recorded on the School Update Form and sent to Westat. Changes in address, principal or school name were recorded on the copy of the PSU List of Schools and sent to Westat.

When the supervisor and school district or private school official had scheduled the introductory meeting, the supervisor completed the Schedule of Introductory Meetings and submitted it to Westat so that Westat could, in turn, send out informational packages and confirmation letters to the appropriate school officials.

7.2.3.2 Introductory Meetings

From August 29 to September 30, 1983, the district supervisors spent about one week in each of their PSUs conducting introductory meetings with school officials. Although the primary purpose of these meetings was to explain NAEP in more detail to the school officials, several other purposes were served as well. During the introductory meeting, the supervisor was responsible for:

- answering questions about NAEP;
- explaining the schools' role in NAEP and distributing the appropriate Summary of School Tasks;
- distributing Student Listing Forms and explaining their use and procedures for filling them out;
- setting up a preliminary schedule for assessments;
- identifying the person within each school who would coordinate all assessment activities;
- collecting and reviewing completed Principal Questionnaires;
- verifying and completing the School Control Form with each principal; and
- obtaining recommendations for exercise administrators, if necessary.

The introductory meetings were the first opportunity for principals and other school officials to discuss the assessment with NAEP staff. Thus, the meetings were particularly important for establishing rapport with the schools, assuring school cooperation, and explaining the details of the schools' tasks to the individuals responsible for them.

7.2.3.3 Schools Added to the Original Sample

Due to a variety of sampling reasons (described in Chapter 4), it was sometimes necessary to add schools to the original sample. Because the process of adding schools to the sample did not begin until late September, when introductory meetings were already taking place, the procedures for contacting and gaining cooperation from these schools necessarily differed from that described for the original sample. For the added schools, Westat first mailed a letter to the district superintendents and heads of private schools. Then, the district supervisor telephoned the contact person in the superintendent's office and asked him or her to notify the sample schools. Westat then mailed a principal's package with Student Listing Forms and the Summary of School Tasks to each school. After three to four days the supervisor called the school and conducted the introductory meeting by telephone. ETS regional staff provided assistance as needed in contacting districts and individual schools. Whenever an in-person introductory meeting was considered essential to insure cooperation, the district supervisor scheduled the meeting during the time that he or she would be in the PSU for the first round of assessments.

7.2.4 Recruit and Train Exercise Administrators

An important part of the supervisors' pre-assessment responsibilities was to hire and train exercise administrators, the persons whose primary function it was to administer the assessment booklets to the sample students. District supervisors were encouraged to use their own discretion in planning for and hiring exercise administrators. Westat provided guidelines for the number of exercise administrators to be hired and also the names of possible exercise administrator candidates located in the supervisor's PSUs.

Supervisors were told that, in general, two exercise administrators should be hired for each PSU, although a variety of factors might influence the actual number. The number of schools in a PSU, the size of the student sample in each school, distances to be traveled, the geography of the area, and weather conditions during particular times of the year were all factors considered by supervisors in developing plans for exercise administrators. A few supervisors had contiguous PSUs; they hired the same exercise administrators to work in all of their PSUs. Other supervisors had PSUs where schools were small and widely scattered; they tended to hire exercise administrators to work only a portion of the PSU.

Candidates for the exercise administrator positions came from several sources. Exercise administrators from previous assessments applied for the jobs. Westat consulted their file of field personnel who had worked on previous Westat studies. Supervisors also used recommendations from school officials for uncovering good candidates. If necessary, advertisements were placed in local newspapers.

Supervisors were encouraged to hire locally, and to hire individuals with teaching experience or the ability to handle classroom situations. Many exercise administrators were retired or substitute teachers.

Training the exercise administrators was one of the supervisor's first tasks upon arriving in the PSU before beginning the assessments. Prior to the supervisor's arrival, Westat sent trainees the Exercise Administrator's Manual which described, in detail, the role of the exercise administrator and procedures to be followed. Exercise administrators were required to study the manual before being trained, then attend a half-day training session conducted by the supervisor. During the training, the supervisor reviewed, in detail, all aspects of the exercise administrator's job, including preparing materials, booklets and administration schedules for assessments; the actual conduct of the session; post-assessment collection of booklets and pencils; coding of booklet covers; recordkeeping; and administrative matters.

7.3 Year 15 Assessments

From October 10, 1983 to May 11, 1984, the assessments were conducted one grade/age at a time. Each supervisor cycled through the four PSUs in his or her region, completing all assessment activities for a grade/age in a PSU before moving on to the next PSU. Ten weeks each were available for supervisors to complete Grade 4/Age 9 and Grade 8/Age 13 assessments; nine weeks were available to complete Grade 11/Age 17. In general, supervisors spent from two to two and one-half weeks for each grade/age class assessment in each PSU.

Supervisors developed their own schedules for each PSU depending upon the size and location of schools, the number of students to be assessed, and any special situations or requests by the schools regarding the timing of sessions. School holidays and requests such as "not on Mondays or Fridays," "only in the mornings," "all students assessed at the same time," etc. were honored by supervisors in arranging the assessment schedule. Such requests affected not only the assessment schedule but also the number of exercise administrators needed at each school.

Although flexibility had to be the hallmark of assessment scheduling, supervisors generally followed the work plan detailed in their field procedures. In essence, this plan involves the following order of supervisory activities upon arriving in a PSU: meet exercise administrators and as part of their training, take them to a school to observe sampling; complete exercise administrator training; draw samples in one or two other schools with exercise administrators; begin assessments in the first school and observe exercise administrators; sample other schools while exercise administrators continue assessments. Where feasible, the supervisors went to each school on assessment day to confirm all arrangements and initiate activity. Depending upon the number of schools in the assessment, the supervisor would schedule sampling and assessments on different days so that he or she could be present at all assessments. Similarly, most supervisors found it very useful to have at least one of the exercise administrators assist with sampling. The supervisor would do the actual sampling while the exercise administrator would double-check the forms, fill out administration schedules, and check the school files for other data, if necessary.

In addition to the activities listed above, the supervisors contacted schools in the next PSU to establish the assessment schedule; called schools in the current PSU to confirm actual assessment dates; made return trips to schools where assessments had been completed to pick up survey forms that had not been finished at the time of assessment; and edited, boxed and shipped completed assessment materials.

7.3.1 Drawing the Sample of Students

Supervisors called each school seven to ten days prior to the assessment to confirm all arrangements for sampling and assessment. The time between sampling and assessment was, on the average, about four days, depending upon the school's time constraints for notifying parents, teachers, and students.

For those Grade 11/Age 17 schools with at least three sessions, supervisors were encouraged to draw the sample during the Grade 4/Age 9 assessment because there is less time available in the spring for Grade 11/Age 17 activities. Also, since high schools tend to be large, sampling in these schools is more time-consuming than in smaller schools. All supervisors tried early sampling; some abandoned it for various reasons. In some schools, because of either geographic location or high rate of turnover in the student body, it did not make sense to attempt to sample early.

Sample selection for Year 15 was more complicated than for previous years for two reasons. First, Year 15 included as eligibles all students in the modal grade[1] as well as those who were age-eligible (the previous NAEP eligibility criterion). Grade eligibles were included for the first time so that the data could be analyzed by grade as well as by age. Supervisors had to check the Student Listing Forms carefully to make sure that only eligibles were included and that all eligibles were included. This proved to be important because, in several instances, supervisors discovered that schools had listed only students who were both age- and grade-eligible. Similarly, supervisors frequently found one or two students who were erroneously listed on the Student Listing Form. This was particularly true in the case of Grade 11/Age 17. The age definition for Grade 11/Age 17 spans two calendar years (October 1966 to September 1967); thus, checking birthdates was more time-consuming because both month and year had to be checked. Also, some of the students sampled as 11th graders during the early sampling in winter had been promoted to the 12th grade at mid-year (before the assessment in spring), modifying their eligibility status. Similarly, 10th graders promoted to the 11th grade had to be added to the Student Listing Form and given a chance of selection.

The second factor complicating sampling was the addition of spiral sessions to the existing practice of tape sessions.[2] Students had to be sampled at different rates for spiral and tape sessions and only age-eligible students were eligible for tape sessions. Thus, supervisors sampled for spiral sessions first. Then, renumbering those age-eligible students who had not been sampled for spiral, supervisors selected the tape session students.

Instructions for sampling were provided in the Supervisor's Manual. Because of the attention to detail required in the sampling, supervisors were required to do all sampling themselves and could not delegate this responsibility to exercise administrators except under extraordinary circumstances which had to be reviewed with the field director.

---

[1] The modal grade is the grade attended by the majority of age-eligible students, that is, the 4th grade for 9-year-olds, the 8th grade for 13-year-olds and the 11th grade for 17-year-olds.

[2] In the tape sessions, all students were administered the same type of booklet and a paced tape was used. In the spiral sessions, each student received one of 63 different self-administered booklets. Chapter 4 provides more information regarding tape and spiral sessions.

Sampling was monitored by Westat statistical staff in several ways, including through the design of the sampling instructions that were sent to supervisors (the Session Assignment Forms). Using school enrollment information on the Principal Questionnaire, the Session Assignment Form for each school provided a range within which the count of names on the Student Listing Form had to fall. If the count of names exceeded either the upper or lower limit, the supervisor had to call Westat. Gross errors in preparing the Student Listing Form could be detected at this stage. For example, if a school included only students in the grade who were age-eligible, the number of names on the Student Listing Form would usually fall below the lower limit on the Session Assignment Form.

In addition, each supervisor was required to report by telephone the following information to the statistical staff for the first school sampled:

(1) PSU;

(2) school ID number;

(3) total of students listed on the Student Listing Form, including any lined out;

(4) total of students lined out on the Student Listing Form;

(5) last line number on the Student Listing Form;

(6) total of students selected for spiral session(s), excluding any lined out;

(7) if tape session in school, the number of age-eligible students (e.g., 13-year-olds), excluding those lined out and sampled for a spiral session; and

(8) total of students selected for tape session(s), excluding any lined out.

Using this information and the sampling rates specified on the Student Listing Form, the statistical staff checked that the sampling had been carried out correctly. The statistical staff also was available to answer questions from the supervisors.

Verifying the sample was also a primary focus of the quality control visits made by Westat and ETS staff. With very few exceptions, supervisors carried out their sampling responsibilities carefully and conscientiously. In the case of one supervisor, it was felt that additional site visits by Westat staff were necessary until satisfactory performance was assured. The details of these visits are discussed in the Report on Sample Selection, Weighting and Variance Estimation: NAEP Year 15 (Lago, Burke, Tepping, & Hansen, 1985).

7.3.2 Conduct of the Assessments

It was, perhaps, in the arrangements for the actual conduct of the assessment sessions that the district supervisors and exercise administrators had to be the most flexible and diplomatic. The physical space and time available in the schools often did not meet the ideal as specified in the manuals. In general, elementary schools were the most flexible and were willing to adapt to the district supervisor's schedule. This was fortunate because the Grade 4/Age 9 students were assessed during the winter and there were times when supervisors had to cancel and reschedule sessions because of inclement weather. The junior and senior high schools were less flexible and were more likely to make special requests for scheduling and timing. For example, some large schools wanted all spiral sessions administered at the same time in an auditorium. To accommodate this, the supervisor acted as an exercise administrator and sometimes had to train the school's teachers in NAEP procedures so they could act as exercise administrators. A session typically ran about one hour, and only one school required that the assessment be done within the time limits of its 40-minute class periods. To accommodate this, the Introduction and Part I of the booklets were administered one day and Parts II-IV were administered the following day.

Another school request which demanded flexibility on the part of supervisors and modified procedures was that all eligible students be assessed, not just those who were sampled. Schools generally made this request when the sample of students to be assessed included all but a few students in a class. In these cases, the school preferred that the whole class be assessed so that the teacher could do other things and the not-in-sample students would not feel that they had been excluded for some reason.

Although supervisors had to be flexible in arranging and staffing sessions, the schedule of activities on the day of assessment was standard. The supervisor and exercise administrators would arrive early at the school to meet with the coordinator and review the assessment plan. The exercise administrators (and supervisor if he or she would be conducting sessions) would assign booklet numbers to the students listed on each Administration Schedule (a listing of the names, ages and sex of every student invited to a session), as specified in the manual. They would then go to the assigned location for the first session and wait for the students to arrive.

Supervisors found it very helpful to have as coordinator someone who was interested in NAEP and willing to make sure that students got to the appropriate sessions. By emphasizing that makeup sessions would have to be scheduled if attendance was low, supervisors were often able to galvanize the coordinators into action to get the students to the appropriate sessions. Actively involved coordinators made sure that teachers and students knew about NAEP, used the public address system to announce sessions and call out the names of missing students, and went from classroom to classroom to hunt for missing students. If the supervisor was not conducting sessions, he or she would do some of these same things and encourage the school staff to make every effort to increase attendance.

Makeup sessions were required for tape sessions whenever attendance at a single tape session was 50 percent or less for the Grade 4/Age 9 and Grade 8/Age 13 schools and 75 percent or less for the Grade 11/Age 17 schools. Makeup sessions were required for spiral sessions whenever attendance at a school's combined sessions was 75 percent or less. Information on makeup sessions is summarized in Table 7(2).

In Year 15, as in previous years of NAEP, makeup sessions were required most frequently during the Grade 11/Age 17 assessment. In fact, makeups were required in less than one percent of the Grade 4/Age 9 and Grade 8/Age 13 schools, but in slightly over 20 percent of the Grade 11/Age 17 schools. Even though the attendance rate requirement was the same for spiral and tape sessions, of the Grade 11/Age 17 schools, about 19 percent of those with spiral sessions required makeups while about 24 percent of those with tape sessions scheduled makeups. This higher makeup rate for tape sessions may have resulted because the attendance requirement applied to each tape session individually but to the spiral sessions combined. Also, a student could attend any of the spiral sessions but, if sampled for tape, he or she had to attend the specified session.

As shown in Table 7(3), makeup sessions represented a small proportion of the number of sessions conducted. Even for Grade 11/Age 17, makeups were less than 10 percent of all sessions conducted (6 percent of all spiral sessions and 20 percent of all tape sessions).

As shown in Table 7(3), a total of 161 makeup sessions were held: 95 spiral and 66 tape. The purpose of makeup sessions was to improve the response (i.e., attendance) rate for Year 15; the actual effect of makeup sessions on response rates is shown in Table 7(4).

Because only six spiral and no tape makeup sessions were required for Grade 4/Age 9 and Grade 8/Age 13, the impact on overall attendance rates was minimal. However, for Grade 11/Age 17, where about 9 percent of all sessions were makeups, there were significant increases in the response rate.

7.3.3 Students Sampled, Invited and Assessed

As mentioned earlier and described in Chapter 4, the combined use of tape and spiral sessions, along with the introduction of grade as well as age samples, complicated the sampling process. A target number of students completing assessments was established for each age group separately for spiral and tape samples. Then, using data from previous assessments on percent excluded and response rates, sample sizes were determined for Year 15. As shown in Table 4(5) (Chapter 4), the actual numbers of students assessed were considerably higher than the target numbers for each grade/age.

Table 7(2)

Frequency of Makeup Sessions

	Number of Schools		
Grade/Age	With Sessions	With Makeup Sessions	Percent with Makeup Sessions
4/9	661	2	0.3
8/13	478	4	0.8
11/17	326	67	20.6
Total	1465	73	5.0

Table 7(3)

Regular and Makeup Sessions Conducted

	Number of Sessions			Number of Makeup Sessions			Percent of Makeup Sessions		
Grade/Age	Spiral	Tape	Total	Spiral	Tape	Total	Spiral	Tape	Total
4/9	1330	260	1590	2	0	2	0.2	0.0	0.1
8/13	1317	261	1578	4	0	4	0.3	0.0	0.3
11/17	1416	324	1740	89	66	155	6.2	20.4	8.9
Total	4063	845	4908	95	66	161	2.3	7.8	3.3

Table 7(4)

Change in Attendance Rates With Makeup Sessions

	Change in Rates (%)	
Grade/Age	Spiral Sessions	Tape Sessions
4/9	+1	0
8/13	+1	0
11/17	+4	+4
Overall	+1	+2

7.3.4 Supervisors' Other Assessment-Related Tasks

A variety of other tasks were undertaken by the district supervisors to assure the successful completion of the assessments and to gather other survey data required by NAEP. Among these supervisory tasks were completing assessment reporting forms; finalizing arrangements for the assessments; supervising exercise administrators; distributing and collecting other data forms and questionnaires; editing, boxing and shipping assessment materials; mailing thank-you letters to coordinators; and filling out a project evaluation for Westat.

When sampling was completed, the supervisor and/or the exercise administrators filled out an Administration Schedule for each assessment session to be held in the school. The administration schedules were the student rosters for the assessment sessions. They identified which students were to attend each session and the time and location of the sessions. Some schools used copies of the Administration Schedules to notify teachers and students. Others wanted an appointment card for each student, which the exercise administrators filled out from the Administration Schedule.

The supervisor also filled out a School Worksheet, containing information on the number of students absent and assessed. Because of the variety of forms and materials pertaining to each school, Westat developed a school folder which could be used by the supervisor to keep all materials pertaining to a school.

7.3.4.1 Finalizing Arrangements for the Assessments

The process of finalizing arrangements for the assessment sessions began prior to the introductory meeting. The supervisor developed a general plan for completing all the assessments in a PSU, taking into consideration each school's geographic location and number of sessions. At the introductory meeting, each school's schedule and constraints were identified and a tentative date established. In general, this date specified the week the assessment would occur, since the supervisor was advised to wait until all meetings had been held and the schedule for all schools known before setting up specific dates with schools. Some schools, however, insisted that the actual dates of assessment be set at the time of the introductory meeting.

At the introductory meeting, the supervisor completed a School Control Form to let the home office know the schedule of assessments. Westat then sent a confirmation memo to the schools and a reminder letter about two weeks prior to the assessment.

Seven to ten days before the assessment week, the district supervisor called the school coordinator to establish (or confirm) the definite dates for sampling and assessment. At the time of sampling, dates and times for the individual sessions were confirmed and recorded on the Administration

Schedules. Depending upon the time lag between sampling and assessment, the supervisor would contact the school coordinator one or more times to confirm arrangements.

Since district supervisors were busy in schools and were hard to reach during the day, schools were instructed to call Westat home office staff if they needed to get in touch with the supervisor. Home office staff received an average of three to five calls per day from schools with questions or requests for schedule changes. If possible, home office staff resolved their questions. If necessary, calls were made to the supervisor's home or hotel, or even to the school where he or she was working.

On the day of an assessment, the supervisor usually went to the school to oversee all assessment activities, handle any special situations that arose and, if necessary, make minor changes in the location or time of sessions.

7.3.4.2 Supervising Exercise Administrators

Supervisors were responsible for the work of their exercise administrators. Because the supervisor was frequently in the school, at least through the first assessments of the day, he or she had ample opportunity to observe the exercise administrators at work. It was mandatory that the supervisors observe the first assessment sessions conducted by each exercise administrator and review the exercise administrator's coding of booklets. Supervisors reported each observation of an exercise administrator on a Weekly Status Report.

District supervisors had the authority to dismiss exercise administrators they considered incompetent and to retrain exercise administrators as necessary. Supervisors took this responsibility seriously; in general, exercise administrators conducted sessions competently and with minimum disruption to the schools.

7.3.4.3 Distributing and Collecting NAEP Questionnaires

The School Characteristics and Policy Questionnaire, Excluded Student Questionnaire and Teacher Questionnaire were distributed in the schools to be completed by school personnel. If these forms were completed in time, the supervisor collected them and shipped them to ETS. The School Characteristics and Policy Questionnaire was mailed by Westat to the school prior to the assessment with the confirmation memo. All other forms were distributed to the school at the time of the assessment by the supervisor.

An Excluded Student Questionnaire was to be filled out for every student who was sampled but was ineligible or excluded from the assessment. The majority of excluded students were those who were determined by the school to be unable to participate in NAEP because they were of limited English-speaking ability, educable mentally retarded, or functionally disabled. For each of these students, the supervisor gave a questionnaire to the coordinator and asked that it be filled out by a teacher of the student. If a student was excluded because he or she was no longer enrolled in the school or had been sampled although ineligible for the study, the supervisor filled out the form. Year 15 is the first year that detailed data have been obtained on the excluded students (see Chapter 6).

The Teacher Questionnaire is also new with Year 15. For this survey, a subsample of students sampled for spiral sessions was identified. The subsample was equal to the number of spiral sessions assigned to the school. Thus, if there were six spiral sessions assigned to a school, a subsample of six students already sampled for spiral sessions would be selected. The school coordinator was asked to identify the English or Language Arts teacher of each student so identified. Those teachers were asked to complete a Teacher Questionnaire.

The supervisor attempted to obtain completed questionnaires from the school by the time he or she completed other assessment activities. If school staff could not give completed forms to the supervisor, the supervisor left an envelope for the coordinator to mail completed forms to ETS.

Initial response for all three questionnaires was very good. Overall, 92.7 percent of the Excluded Student Questionnaires, 88.7 percent of the School Characteristics and Policy Questionnaires and 86.9 percent of the Teacher Questionnaires that had been distributed were collected and returned by the supervisors to ETS. Response, although high, was lowest for the teacher survey. This may have been because the questionnaires were passed from the supervisor to the school coordinator to the teachers, creating greater opportunity for the questionnaires to be misplaced and greater difficulty in collecting completed questionnaires.

7.3.4.4 Editing, Boxing, and Shipping Assessment Materials

Selected items from the Administration Schedule were coded onto the front cover of the assessment booklets. This responsibility was shared by supervisors and exercise administrators, although supervisors had to review all work completed by the exercise administrators. Supervisors were to ship to ETS all assessment materials for a school within a week of completing the assessment in that school. At the end of assessments in each PSU, supervisors shipped PSU-specific materials to ETS; at the end of assessments for a grade/age, the tapes and forms specific to that grade/age were shipped back. At the end of the field period, all materials were either returned to ETS, returned to Westat, or discarded.

For materials returned to ETS, district supervisors completed and mailed separately to ETS a pre-printed postcard upon which they recorded the shipment date, PSU number, school number, number of cartons shipped and the mode of shipment (U.S. mail, United Parcel, etc.).

If after seven days from receipt of the mail-alert postcard the materials had not been received at ETS, Westat was notified. Westat in turn contacted the district supervisor to begin the process of tracing the shipment. Fifty-five assessment books were lost or damaged in transit.

7.3.4.5 Close-Out Activities

At the end of the field period, district supervisors were sent copies of a thank-you letter to the school coordinators. This letter was signed and sent by the supervisor as a personal thanks to the coordinator. At the same time, Westat mailed letters of appreciation to superintendents and heads of private schools. School principals were sent a certificate of appreciation from ETS.

As a final task, district supervisors were asked to complete and return to Westat an evaluation of Year 15 field activities. The recommendations made by the supervisors will be incorporated into future assessments.

7.3.5 Quality Control and Evaluation Studies

There were two specifically designed quality control studies of the field effort. The first, and most intensive, involved on-site visits by Westat and ETS staff to verify the sampling and to observe the supervisors and exercise administrators as they conducted assessments. The second study was a telephone survey of a 10 percent sample of schools. This survey took place after the field period had ended and all assessment activities had been completed in the schools. As part of the telephone survey, the school coordinators were thanked for the school's participation and asked about their experiences with NAEP and the field staff.

7.3.5.1 On-Site Quality Control Visits

At the beginning of each grade/age assessment, a sample of schools was selected for quality control visits by Westat and ETS staffs. The purpose of these visits was twofold: first, they provided data from which rough estimates could be made of the quality of assessment activities, particularly the sample selection of students. The second purpose of the quality control visits was to observe, in person, the work of supervisors and exercise administrators to identify and correct areas of confusion or error.

The design of the sample of schools for the quality control visits sought to satisfy both purposes of the visits through a combination of purposive and probability sampling. The probability sample was designed to provide data to assess the quality of assessment activities. The purposive sample (where Westat field management specified which supervisors should be visited) allowed judgmental selection of those supervisors who we thought could benefit from further observation.

The number and distribution of quality control visits among grade/ages is provided in Table 7(5).

Because of the importance of identifying problem areas and taking corrective action as quickly as possible, half of all quality control visits (two thirds of the purposive visits) were scheduled from October to December 1983, when the Grade 8/Age 13 assessments were taking place. Thirty-two schools were visited during that period, sixteen by Westat and sixteen by ETS. Each supervisor was visited twice, once by ETS and once by Westat.

During the next grade/age assessment, Grade 4/Age 9, visits were made to schools of twelve of the sixteen supervisors. Schools in eight of the supervisory regions were selected at random; the remaining four were selected purposively. In anticipation of special problems in high schools, given their size and relatively lower attendance rates in previous years of NAEP, the number of quality control visits during this third grade/age was increased to 20 so that each supervisor's region was visited at least once, and four were visited more than once.

In general, the visits went well. (The Report on Sample Selection provides more specific results of the quality control visits.) The sampling problems that were identified tended to be occasional random errors due to carelessness rather than systematic errors reflecting a misunderstanding of procedures. Similarly, the kinds of procedural mistakes made by supervisors and exercise administrators tended to be stylistic (not speaking loud enough, not following the prescribed script) rather than a result of misunderstanding. These issues were discussed with the individual supervisor as they occurred. If applicable, a general field memorandum was prepared on specific issues and distributed to all supervisors.

7.3.5.2 Telephone Survey

In early May, Westat selected a 10 percent sample of the participating schools for inclusion in a telephone survey. The purpose of the telephone survey was to give the principals or school coordinators an opportunity to comment and make suggestions about operational procedures and the conduct of the field staff. Details concerning the telephone survey and questionnaire are contained in the Report on Field Operations.

7.4 Field Management

Various administrative reporting forms were developed for pre-assessment and assessment activities in order to monitor the progress of work. These forms and how they were used are described in this section. Copies of these forms can be found in the Report on Field Operations.

Table 7(5)

Number and Distribution of Quality Control Visits

	Number of Schools in Sample		
Grade/Age	Purposive	Probability	Both
8/13	16	16	32
4/9	4	8	12
11/17	4	16	20
Total	24	40	64

7.4.1 Monitoring Field Activities

Several approaches were taken to monitor the progress of work during the pre-assessment and assessment phases.

During the pre-assessment activities (arrangements for and conduct of introductory meetings), the district supervisors reported to the field director at Westat at least once a week to review progress in scheduling introductory meetings and to discuss any problems or difficulties. Each week the receipt clerk reported to the field director the number of Results of Contact Forms received at Westat. This report provided a good indication of the progress of the contacts and scheduled meetings. In addition, an automated management system was developed which contained a record for each sampled school. A disposition code structure was developed to indicate the status of the school's participation (e.g., cooperating, school refusal, district refusal, school closed, school dropped--no age eligibles, etc.). When a Schedule of Introductory Meetings was received at Westat, the receipt clerk keyed a cooperating disposition code for each school invited to attend the introductory meeting. If a school or school district refused, as noted on the Results of Contact Form, a refusing disposition code was keyed for each refusing school. The School Update Form was the source of disposition status for schools that closed, had a grade or enrollment change that made them ineligible, or had no age-eligible students.

Disposition reports were generated from the receipt system at least once a week to review the progress of securing cooperation from the sampled schools. Four different reports were generated during the pre-assessment activities. The first report gave a breakdown of the number of schools by disposition code in each PSU. The second report listed the ID number and school name for each cooperating school for which a Principal Questionnaire had not been received. The third report listed the ID number and school name for each school without a disposition code. The fourth report listed the ID numbers and school names for non-cooperating schools (refusals, school closed, no age eligibles) and their disposition codes.

These reports were an invaluable tool for the sampling statisticians as well as for the field director and assistant field director. They provided the statisticians with the information needed to determine whether the sample of schools was adequate to produce representative results. Based on the information contained in these reports, the sampling statisticians substituted new schools into the sample to replace some of the non-cooperating schools and supplemented the original sample with additional schools where needed.

During the assessment activities the automated management system was expanded to include the results of the actual assessments. Data from the School Worksheet on number of students to be assessed, number assessed, and number absent were keyed by session for both spiral and tape sessions. In addition, data from the Roster of Questionnaires on the number of Excluded Student Questionnaires and Teacher Questionnaires expected and shipped to

ETS, and whether the School Characteristics and Policy Questionnaire was shipped to ETS were also keyed into the receipt system.

A response rate report was generated weekly which allowed the project staff to monitor the progress of the assessments by checking both that the schools were assessed on schedule and that a high response rate was achieved. The sampling statisticians used these reports to monitor the sample yield by school, PSU and grade/age.

Another method used to monitor the progress of the assessments was the twice-weekly telephone report between the field director or assistant field director and each supervisor. During these phone conversations, the field director and assistant field director reviewed the supervisor's schedule as well as any problems the supervisor was experiencing.

The district supervisors were required to complete a monthly calendar which indicated when each school was being sampled and assessed. This calendar served two purposes. First, it allowed the field director and assistant field director to review the supervisor's schedule and the distribution of work for the month. Second, it enabled the field director and assistant field director to locate the supervisor when urgent messages had to be relayed.

In addition to the monthly calendars, the supervisors completed a Weekly Status Report. This report indicated which schools had been called to confirm the assessment date(s) as well as the dates that the sample was drawn, assessment completed, school shipment mailed to ETS, and when exercise administrators were observed for each school. Information from this report was reviewed with the supervisor during the twice-weekly telephone call.

7.4.2 Materials and Reporting Forms for Assessment Activities

At the second training session in October (prior to the start of assessment activities), Westat provided the supervisors with the reporting forms and supplies needed for the assessment phase, as well as the Session Assignment Forms which were used to sample students for the assessment and sample teachers for the teacher survey. ETS also provided materials and supplies necessary for the conduct of the assessment and shipped them to the supervisors' homes. At the start of each grade/age assessment, additional materials were shipped to replenish supplies.

Supplies provided by Westat included No. 2 pencils, pencil sharpeners, appointment cards, timers, tape recorders, additional supplies of parental consent letters and NAEP brochures, and Dali postcards for assessment (the postcard was placed inside specific test booklets and was used by the students assigned these booklets to complete some of the test items).

During training, the supervisors were given a binder with a Session Assignment Form for each sampled school in the Grade 8/Age 13 sample (Grade 8/Age 13 assessments began the week after training). The session

Assignment Form provided the supervisor with the number of each type of session (spiral and tape) to be held in the school as well as the line numbers that designated the students selected for assessment from those listed on the Student Listing Form. In addition, line numbers were provided for use in selecting teachers for the teacher survey. Session Assignment Forms for the other two grade/age assessments were sent several weeks prior to the start of each grade/age assessment.

When the student sample selection had been completed and the selected students had been assigned to sessions on the Student Listing Form, the supervisor completed an Administration Schedule for each session to be held in the school. These Administration Schedules served as student rosters to be used by the school coordinators and exercise administrators to carry out the sessions.

On each Administration Schedule, the supervisor entered the day, date, time, and location of the session, type of session (spiral or tape) and the name and ID number of the exercise administrator conducting the session. The supervisor then carefully transferred the name, homeroom, birthdate, grade, and sex of each student assigned to that session from the Student Listing Form. The Administration Schedule used in the Grade 4/Age 9 and Grade 8/Age 13 schools had two copies and the schedule used in Grade 11/Age 17 schools had three copies. The additional copy was added for the Grade 11/Age 17 schools since the second copy (which was retained by the school) was usually given to the school coordinator prior to the start of a session and therefore did not have booklet numbers recorded on it. (The school needed to retain a copy of the Administration Schedule with student booklet numbers in order to participate in a follow-up language study.)

The supervisor gave the top copy (or top two copies for the Grade 11/Age 17 assessment) of the Administration Schedule to the exercise administrator who was to conduct the session. The exercise administrator used this copy during the session to check attendance, observe race/ethnicity and record the student identification number from the assessment booklet. After the session, the exercise administrator reported the results of the session by entering the number of students assessed, number of students absent, bundle numbers from which booklets were used, and number of used and unused booklets. The exercise administrator then tore off the top copy at the perforation (between "Homeroom" and "Birthdate"). The portion of the top copy of the form containing the names was left with the school coordinator (who also retained the second copy of the entire schedule) and the tear-off portion without the names was given to the district supervisor. The district supervisor and exercise administrators used the tear-off portion to code the front cover of the test booklets and mailed this tear-off portion to ETS with the test booklets and other reporting forms.

The Roster of Questionnaires served as an important recordkeeping and shipping document for Excluded Student Questionnaires and Teacher Questionnaires. This form was printed on three-part paper and completed by the district supervisor. One copy of the form was sent to ETS, one copy was sent to Westat, and one copy was retained by the supervisor.

The first section of the Roster of Questionnaires contained relevant information about the excluded students. The supervisor recorded the excluded student's line number from the Student Listing Form, the type of session for which the excluded student had been selected, the ID number from the Excluded Student Questionnaire, and information about shipment of the Excluded Student Questionnaire to ETS.

The second section of the Roster of Questionnaires pertained to the teacher survey. Here the supervisor recorded the line number from the Student Listing Form of the student selected for the teacher survey, the code number assigned to the teacher, the ID number from the Teacher Questionnaire, and information about shipment of the Teacher Questionnaire to ETS. Also, at the bottom of the Roster was a place to indicate whether the School Characteristics and Policy Questionnaire was enclosed with the shipment of school materials to ETS.

If the district supervisor was unable to collect all of the Excluded Student Questionnaires, Teacher Questionnaires, and/or the School Characteristics and Policy Questionnaire before shipping the assessment materials for a school to ETS, he or she made follow-up contacts with the school to collect the remaining questionnaires. When remaining questionnaires were collected and shipped to ETS, the supervisor completed a Supplemental Transmittal Sheet which contained the ID numbers of the questionnaires included in the shipment as well as the ID numbers of questionnaires that had not yet been transmitted. This form was also printed on three part paper; one copy was sent to ETS, one copy was sent to Westat, and one copy was retained by the supervisor.

The School Worksheet was the control document used by the district supervisor to report the results of the assessment in a school. It was printed on three-part paper; one copy was sent to ETS, one copy was sent to Westat, and one copy was retained by the supervisor. ETS used this document to verify the supervisor's shipment of the school's assessment materials. Westat used this document to enter the results of the assessment (number of students to be assessed, number assessed, number absent, session number, assessment date and room in which the assessment was conducted) into an automated management system.

Session attendance results and bundles used were entered by session for spiral and for tape on the School Worksheet. At the bottom of the form, the response rate for all spiral sessions and the response rate for each tape session was calculated in order to determine the need for a makeup session. If a makeup session was necessary, the results of the makeup session were recorded on the School Worksheet for Makeup Sessions.

An Exercise Observation Sheet was completed by the district supervisor whenever he or she or an exercise administrator noticed a problem related to an exercise. These sheets were mailed with the school's assessment materials to ETS.

A Session Header Form was completed for each session in a school and placed on top of the booklets used in the session when the school's assessment materials were shipped to ETS. The Session Header Form contained the supervisor's name, the PSU and school numbers, and the session number.

ETS was responsible for the distribution of materials relating to the actual assessments, such as test booklets and questionnaires, as well as supplies used for the shipment of assessment materials. At the start of work in each PSU, ETS shipped these materials and supplies to the district supervisors.

With the materials and supplies for the first PSU in a grade/age assessment, ETS sent two sets of each type of stimulus tape to be used in the tape sessions.

7.4.3 Materials and Reporting Forms for Pre-Assessment

In August 1983, ETS regional office staff made the initial contact with public school superintendents and private school officials. The Results of Contact with Superintendent/Private School Official Form was designed to document the results of these calls. When a final determination had been made regarding whether or not the school would participate, completed forms were mailed directly to the appropriate district supervisor by the ETS regional office. The form has two parts--Part I was completed by ETS staff and Part II was completed later by the district supervisor. The form is printed on four-part paper; one copy was retained by ETS, two copies were sent to the district supervisor, and one copy was mailed to Westat. Once the supervisor received this form from ETS, the supervisor proceeded to call the school to arrange for an introductory meeting. These arrangements were entered on Part II of the form.

Using the information in Part II of the Results of Contact Form, the district supervisor completed the Schedule of Introductory Meetings. On this form the district supervisor entered the date, time, and location of the introductory meeting and listed the names of all persons asked to attend. This form was mailed to Westat with one copy of the Results of Contact Form. Using the Schedules of Introductory Meetings, mailing clerks at Westat prepared letters specifying the date, time, and location of the introductory meeting and mailed them to each of the persons entered on the schedules.

During the telephone call by the district supervisor to the public school superintendent or private school official to set up the introductory meeting, the supervisor reviewed and updated his or her copies of the PSU Listing of Selected Schools. The update was intended to uncover new schools that may have opened, sampled schools that may have closed, changes in grade span or enrollment, or corrections in the name and/or address of the superintendent, principal or school. The district supervisor was given two copies of the PSU Listing of Selected Schools. All name and address changes were recorded on the two copies of this listing and one copy of the listing was mailed to Westat. All other types of changes (e.g., school openings or closings, or changes in grade span or enrollment) were recorded on the School Update Form. The School Update Form was also mailed to Westat and given to the sample statisticians who used it to make adjustments to the sample when necessary.

During the introductory meeting, the district supervisors collected the Principal Questionnaire. The supervisor reviewed the information entered on the Principal Questionnaire for completeness at the meeting and sent completed Principal Questionnaires to Westat.

Prior to the introductory meeting, the district supervisor received two copies of the School Control Form, a computer-generated form containing summary information about the school. The first section of the form, School Information Provided, supplied the supervisor with the estimated total eligible students and the preliminary number of sessions expected in the school. The second part of the form was completed by the supervisor at the introductory meeting. Items completed were the name of the school coordinator, if one had been appointed; the dates of the assessment week agreed upon with the principal; how the school planned to complete the Student Listing Form; and any other information learned about the school's requirements for conducting the assessment. One copy of the School Control Form was mailed to Westat; the second copy was retained by the supervisor.

Chapter 8

MATERIALS PROCESSING AND DATABASE CREATION[1]

John L. Barone

Educational Testing Service

The previous chapter on field administration described the conduct of the NAEP assessment in the field to the point of shipment of materials to ETS. This chapter details the receipt, processing and final disposition of these assessment materials at ETS as they were transcribed to computer-readable form and placed in an integrated NAEP database to be used for data analysis and reporting. This database is now available to external users via the public-use data tapes (PUDTs).

The flow of materials, creation of data files, and creation of the NAEP database are depicted as an ordered set of processes that are applied either to the assessment materials or to the transcribed data. The following chapters describe each of these processes in detail.

The large volume of collected data and the complexity of the Year 15 NAEP design, with its spiralled distribution of many books, required the development and use of NAEP-specific data entry and management systems, including carefully planned and well-defined editing and quality control procedures. This chapter discusses the implementation and use of systems and processes that resulted in data management procedures that were effective, responsive, and insured the quality and integrity of NAEP data. The result is the final NAEP database, which met the original objectives of integrity and usefulness, and exceeded stringent standards for "correctness" and quality.

Figure 8-1 is a flow diagram that shows the conceptual framework of ordered processes that were applied to the NAEP materials and data files. The dashed line through the center of the figure divides the outline into two sets of processes, Processing Assessment Materials and Database Creation, described below.

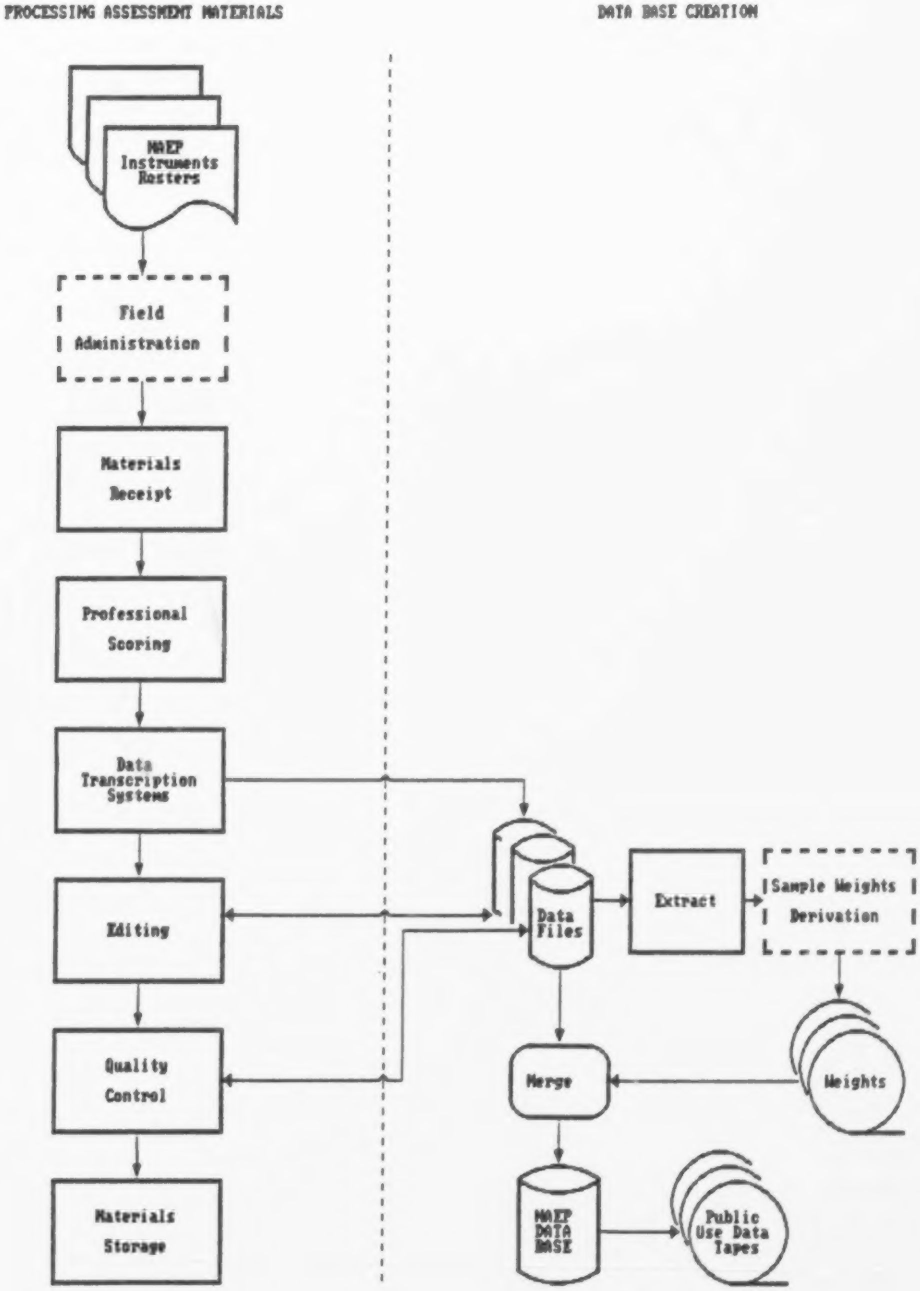
The processes represented by solid-line boxes in the flow diagram were performed at ETS on the paper materials or computer files. The two processes enclosed in dashed-line boxes (Field Administration and Derive Sampling Weights) were performed by Westat and are discussed in detail in Chapters 7 and 4, respectively.

---

[1] Flow diagrams for this chapter were produced by William Van Hassel.

Figure 8-1

Data Flow Overview


PROCESSING ASSESSMENT MATERIALS
DATA BASE CREATION
NAEP Instruments Rosters
Field Administration
Materials Receipt
Professional Scoring
Data Transcription Systems
Editing
Quality Control
Materials Storage
Data Files
Extract
Sample Weights Derivation
Weights
Merge
NAEP DATA BASE
Public Use Data Tapes

8.0.1 Processing Assessment Materials

The left side of Figure 8-1 depicts the flow of NAEP "paper" materials. Chapter 8.1 describes this flow in detail and discusses how information contained on the field rosters, schedules, and worksheets were used as controlling mechanisms for processing of materials. It also follows the path of each assessment instrument (student assessment books, School Characteristics and Policy Questionnaires, Teacher Questionnaires, Excluded Student Questionnaires), school worksheets, and administration schedules as they are tracked through the appropriate processes that result in the final integrated NAEP database.

The following is a brief description of each process involved in materials processing as shown in Figure 8-1. Each description refers the reader to the section(s) or chapter(s) in which the process is discussed in detail.

Field Administration is the conduct and monitoring of the NAEP assessment in the schools. Chapter 7 discusses this process in detail.

Materials Receipt refers to receipt and processing of assessment materials at ETS. Section 8.1.1 describes the procedures and forms that were used to check and verify the receipt of documents from the field. It also discusses the follow-up procedures that were initiated when discrepancies were identified. As a result of this process, paper materials were received and subsequently batched for NAEP materials processing and data transcription.

Professional Scoring is the process that resulted in the scoring of the open-ended NAEP reading and writing items. Chapter 8.2 describes the items, types of scoring used, scoring operation, reliability checks, and resolution of scoring discrepancies. Entry and editing of this data are discussed in Sections 8.1.4 and 8.4.2.

Data Transcription Systems refers to the methodology used to transcribe NAEP materials to computer-readable form. The transcription method used for each NAEP instrument is discussed in Chapter 8.1. Chapter 8.3 describes the design, structure, and development of the NAEP-specific data entry system used to transcribe most of the NAEP materials to computer files; it also discusses the tracking and audit mechanisms that were built into the system to ensure that all data was properly processed and accounted for.

Editing refers to the ETS procedures that ensured the correctness and integrity of the NAEP data files by (1) validating every field of NAEP data that was entered into computer-readable form, (2) identifying any invalid or inconsistent values, and (3) correcting or flagging as unresolvable those values identified as invalid or inconsistent. Chapter 8.4 describes these procedures.

Quality Control refers to the ETS procedures that assessed the accuracy of the data transcription and editing operations. Chapter 8.5 discusses the quality control procedures used in NAEP and provides a summary of the likely error rates.

Materials Storage refers to the final disposition of NAEP "paper" materials after processing had been completed. Chapter 8.1 discusses materials storage.

8.0.2 Database Creation

The right side of Figure 8-1 depicts the evolution of the integrated NAEP database from the transcribed data to the final database, available to external users via the PUDTs. Chapter 8.6 describes the processes through which the database evolved.

The remainder of this section contains a brief description of each process involved in Database Creation as shown in the figure. Each description also refers the reader to the section(s) or chapter(s) in which the process is discussed in detail.

Data Files refers to (1) the data files created by the ETS/NAEP data transcription, editing and resolution systems and (2) the labeling files (discussed in Chapter 8.6) that contain descriptive information on every item used in NAEP.

Extract is the process discussed in Section 8.6.1 that created data files containing specific demographic data fields from the ETS/NAEP data files. These data files were required by Westat to derive sampling weights.

Sample Weights Derivation was performed by Westat and is discussed in Chapter 4. This process produced computer tape files containing sampling weights for every student and school assessed by NAEP.

Merge refers to the final integration of NAEP data files into the NAEP database. This process, discussed in Section 8.6.2, merged the NAEP data files, labeling files, and the NAEP sampling weights into one inclusive database.

NAEP Database is the final, integrated NAEP database that contains all Year 15 NAEP data. This is the database that is ultimately made available to external users via the PUDTs. The structure of the NAEP database is discussed in Chapter 8.6; the PUDTs are discussed in Chapter 8.7.

Chapter 8.1

PROCESSING ASSESSMENT MATERIALS

Alfred M. Rogers
Norma A. Norris

Educational Testing Service

Chapter 7, Field Administration, traced the progress of the assessment booklets and related documents in the field to the point of shipment to ETS. This chapter details the receipt and processing of these assessment materials at ETS.

8.1.1 Materials Receipt

It was the responsibility of the district supervisor to complete and mail a postcard to ETS at the completion of assessment administration in each school. This card contained the assessed school identification, the number of boxes shipped, and the mode of shipment. The receipt of this card at ETS alerted staff to expect arrival of the shipment within seven working days. If after seven days the shipment had not arrived, ETS notified Westat, who in turn initiated a trace of the shipment. This tracing process was successful in all cases except one, in which the full set of assessment materials from one school was never recovered. Some other shipments broke open in transit. In all, 55 booklets were lost or damaged.

The shipment from each school contained the school worksheet; administration schedule; questionnaire roster; School, Teacher, and Excluded Student Questionnaires; and assessment booklets, bundled by session, with session header sheets. The format and content of these instruments are documented in the chapter on field administration. The following discussion of check-in procedures presumes an understanding of information contained in and inter-relationships among these instruments.

The school worksheet contained summary counts of the booklets used in all assessment sessions in each school. The session numbers listed on the worksheet were first checked against the session numbers written on the session header sheets enclosed with each bundle of assessment booklets. The booklets within each session were then counted and checked against both the count written on the session header sheet and the counts of used and unused booklets in the corresponding columns of the school worksheet. All discrepancies in the counts were referred to the administration schedules for resolution. The booklet numbers from the bundle in question were compared against the listing of booklet numbers on the schedule. If the discrepancy could not be resolved by this process, Westat was notified, who in turn contacted the appropriate district supervisor for resolution.

The Teacher and Excluded Student Questionnaires were then counted and compared against the questionnaire roster. Any discrepancy in the Excluded Student Questionnaire counts was referred to Westat and again, in turn, to the district supervisor for resolution. Since the field administration procedures permitted a separate shipment of teacher and school questionnaires, any discrepancy in the Teacher Questionnaire counts alerted the receiving staff to expect a later shipment.

When all of the student-related materials for a school had been received and checked in, the assessment schedules, school worksheet, assessment booklets, and questionnaires were forwarded to the data operations coordinator for transcription processing. The operations coordinator separated these materials according to the appropriate data entry procedures: the assessment schedules were accumulated and shipped in batches to key entry; the school worksheet and assessment session bundles were sent directly to data entry systems; the Excluded Student Questionnaires were also batched and sent to data entry systems as scheduling permitted; and the Teacher and School questionnaires were accumulated and held for data entry until the student and excluded student instruments were completed. The remainder of this section follows these instruments through entry, editing, and quality control processing.

8.1.2 Administration Schedules

As described in Chapter 7, the administration schedules contain the demographic characteristics of the students selected for the assessment. This information, which included the sex, ethnic origin, grade, and birth date of the sampled students, was used by Westat in the derivation of sampling weights. The booklet numbers of the students who participated were transcribed to the schedule at the time of the assessment, and the demographic information was in turn transcribed to the front covers of the booklets after the assessment.

The demographics of the students who were sampled but did not participate in the assessment (exclusions and absentees) were used to adjust the sampling weights of those who did. The excluded student information could be obtained from the Excluded Student Questionnaire data, but the information on absentees could only be found on the administration schedules. It was imperative, therefore, that this information be transcribed to computer-readable media and combined with the assessed and excluded student data.

The administration schedule data was transcribed to computer tape by the key entry systems at ETS. One record was generated for each absent student (line) on the form. The PSU, school, and session codes from the top of the form were repeated for each student on the form. The information transcribed for each absent student included sex, grade, and birth date. These data were ultimately used by Westat to adjust the sample weights.

At the completion of entry processing, the data tape was copied to disk for editing and quality control processing. The editing process consisted of a validation program and an interactive text editor for correcting erroneous data. The validation program checked that the demographic information was present and within the appropriate ranges. The schedules were used in this process to resolve any errors or discrepancies uncovered by the program and to "spot-check" records for quality control.

The assessment schedules were retained by the operations coordinator in anticipation of future questions about and references to the sample. This proved to be the most efficient and compact means of retaining the relevant raw data since the schedules for all three grade/age assessments could be contained in one storage box.

8.1.3 School Worksheets

The school worksheets were forwarded by the operations coordinator to entry staff for processing under the NAEP data entry system. This system was designed and developed by ETS staff to meet the singular requirements of the NAEP Year 15 design, and is more fully described in Chapter 8.3.

Each column of the school worksheet contained information pertaining to the administration activity of each session within a school. This information included the date, time, and location of the administration, the exercise administrator code, and the counts of the students sampled, excluded, absent, and assessed. These data, along with the PSU, school, and session codes, were keyed into the system by entry staff.

To enter this information, entry staff had to first log on to the computer and start the data entry program. The program prompted for the operator's initials, which would be used in subsequent reporting of entry processing activity. The operator was then presented with a primary menu, requesting input of the codes for the instrument to be processed and the processing mode. The codes and their associated actions were listed below the corresponding entry fields. The operator typed in the codes for "School Worksheet" processing under "Entry" mode and pressed the ENTER key. A second screen appeared, requesting input of the PSU and school codes, and the number of spiral and tape sessions to be entered for that school. The operator keyed in these values and pressed ENTER again. The program then presented one entry screen for each session to be entered, automatically assigning the session code for spiral sessions and requesting the booklet number for the tape session code. The operator then keyed in each column of information from the worksheet and pressed ENTER to proceed to the next session. When all sessions for a school had been entered, the program would re-display the school screen if there were more worksheets to process. If the operator had no more worksheets to enter, pressing ENTER with no data in the PSU and school fields would return the program to the primary menu, from which control could be passed to other parts of the entry system.

The entry system controlled the processing of student data and maintained statistics on the entry activity at the session level. This was accomplished by means of a tracking file, on which each record contained all control and reporting information for one session. The entry of the school worksheet information thus generated a new record on the tracking file for each session, initializing the control parameters. The system would not allow entry of student data to proceed unless the school worksheet information had first been entered.

The operations coordinator was provided with procedures for periodically monitoring and reporting data entry activity. These procedures compared the counts of booklets processed at each stage with the initial counts from the worksheet, and flag discrepancies. This, in turn, alerted the coordinator to possible missing or extra booklets. If the school worksheet information was determined to be in error, the operations coordinator had the facility to correct the tracking file data to prevent reoccurrence of the discrepancies in the activity report.

The school worksheets were retained by the operations coordinator in anticipation of later queries, since they could be compactly stored and easily referenced.

8.1.4 Student Assessment Instruments

The student assessment booklets were forwarded directly to the data entry area as the complete set of materials was received from each school. The booklets were bundled by session, with a session header sheet attached to the top of each bundle. This sheet contained the PSU, school, and session codes, serving to identify each bundle. The header sheets were retained with the bundles throughout entry processing.

8.1.4.1 Response Data Entry

The entry operator initiated student data entry by entering the "Student Data" and "Entry" codes on the primary menu. The entry program then displayed a screen requesting input of the PSU, school, and session codes for the session data to be entered. If the tracking file information indicated that entry processing had terminated for that session, the program displayed a message to the operator; and if the session code was correct, the problem was referred to the operations coordinator for correction.

If entry processing was permitted, the program displayed a screen for the entry of student booklet cover information and requested the entry of the booklet serial number. If the booklet number was incorrect, or a booklet with that serial number had already been entered, processing stopped, a message was issued, and the operator could either enter the correct serial number if it was mis-keyed, or set the booklet aside for resolution by the operations coordinator. If the serial number was acceptable, the program prompted for entry of the block codes printed on the cover, to verify that the correct booklet number had been entered. If the block codes had been entered correctly from the booklet but did not agree with the programmed codes, program control was returned to the entry of the serial number and the operator had to again either enter the correct serial number or set the booklet aside for resolution.

On successful entry of the block codes, the program prompted for entry of the remaining booklet cover data fields. This information included the administration code, exercise administrator code, student's sex, ethnicity, grade, and birth date, and the PSU and school codes. The entry of these data followed the same model as the entry of the response data contained in the rest of the student assessment instruments as well as the school, teacher, and excluded student questionnaires. This model will be described below from the viewpoint of the entry operator. An explanation of the program functioning is found in Section 8.3.7.

With the exception of the non-scorable open-ended response items, the responses to all items could be entered from the numeric keypad on the computer terminal keyboards. For the multiple-choice items, the program software automatically converted the entered numeric values into their alphabetic counterparts: "A" for "1", "B" for "2", etc. Three extra keys on this keypad were reserved for special processing codes: the hyphen was entered under a "no response" condition; the period indicated a multiple response to an item where only one response was expected; the comma was converted to a question mark by the program and flagged data which the operator could not resolve immediately and which needed resolution by the operations coordinator or designated entry staff. Additionally, three function keys allowed the operator to control field processing: the TAB key passed control to the next field on the form; the BACKSPACE key passed control to the previous field; and the ENTER key signalled to the program the completion of processing for a field or an entire form.

The program controlled processing of the entered data virtually at the keystroke level, interrupting and alerting the operator only when the data values failed to meet range validation criteria. If an invalid data value was entered, the program "locked" on the problem field, disabling the function keys until a legitimate value was entered. At the completion of the last field on a form, the operator would press ENTER and the program would scan the entered data for blanks, to ensure that no fields had been skipped or otherwise erased. If a blank was found, the operator was alerted and instructed to fill in the problem field.

The open-ended non-scorable items were included in the entry process in an effort to capture all response data. These responses were found in the few items which had a "Specify Other" category with a space to be filled in by the respondent. Only eight characters were permitted by the system for the entry of this information, so operators had to abbreviate or use key words at their own discretion. Those items which requested information on language usage, country or state lived in were codified. The entry system automatically displayed the possible responses and their code values when these fields were encountered in the entry process.

Upon successful entry of the booklet cover information, the program displayed entry screens for each section of the current booklet. The first screen was always for entry of the common background information, since this was the first section in all student booklets. The BIB spiral booklets contained three additional sections; the UBIB spiral and tape booklets contained only two. The type and order of these sections was completely controlled by the booklet number, according to the NAEP design. At the completion of entry for the last section in each booklet, the program re-displayed the booklet cover entry screen to accept input for another booklet. A blank field entered for the booklet serial number indicated the end of entry processing for that session. The program performed session clean-up and re-displayed the session header entry screen in anticipation of entry processing for another session. A blank field entered for the PSU indicated termination of student data entry processing and the program returned to the primary menu.

Several of the participating schools conducted all of their spiral sessions as one large session. Consequently, some session bundles were too large to be accommodated in one entry sitting. The program permitted interruption of entry and verification processing to adapt to the entry operators' schedules. At the completion of entry processing for a session, the operator's initials and the date were written on the session header sheet and the bundle was placed in the staging area for verification processing.

The entry mode created the student data records and wrote them to the entry system work files. The verification mode was essentially a second entry of the data and a blind field-by-field comparison with the original data. If, for any field, the data value entered under verification differed from the initial value, the program would "lock" on that field, issue a message to the operator, and allow the operator to determine whether the value was mis-keyed or incorrectly entered the first time, enter the "correct" value, if necessary, and press ENTER to continue processing the remaining fields. While the program was locked on the discrepancy, the operator could press the question mark key to view the initial data value. During verification processing, each data record was rewritten to the work file with all changed data values.

At the completion of verification processing for a session, the program printed an audit trail listing at a printer in the entry area. This listing was a formatted summary of an adjunct file to the work data file which was created and updated by the system during processing of the session data. A record was written to the audit file whenever the multiple response code or a question mark was entered as a data value under any processing mode, or if, under verification mode, a data value was changed from its original value. Each audit record contained identification information, including the PSU, school, session, booklet serial, section and item numbers of the data value, and the operator code, processing mode, date and time of the action, as well as the old and new data values.

This audit listing was attached to the session bundle and forwarded to the resolution area. Staff assigned to resolution processing reviewed the audit listing, checked the actual responses in the booklets wherever question marks were indicated, determined the appropriate value(s) to be coded in the data file, and wrote these new codes on the audit listing.

The resolution mode of the entry system permitted the operator to access data records, display the field values, and make corrections to individual fields. A change in any data field under resolution mode also generated a record for the audit file, and the program produced a second audit listing at the completion of resolution processing for each batch. There was no limit to the number of times a session or data record could be processed under resolution.

On completion of resolution processing, each bundle was stored in a labeled box and held for final editing and quality control processing.

The final editing was performed after the entry work files had been spooled into a master student data file. This spooling program checked every data field of every student record for out-of-range values and question marks. A listing similar to the audit listings for each session was produced, which resolution staff then used to identify and correct the remaining data anomalies.

The quality control process selected a random sample of each booklet type from the master student file, identifying those booklets for extraction from the raw data. The designated booklets were located, pulled from their boxes, and forwarded to quality control staff. The responses in each booklet were then compared with their coded data values in the data file. The full details and results of the quality control process are presented in Chapter 8.5. On completion of quality control processing, the booklets were returned to their boxes and shipped to the professional scoring area.

8.1.4.2 Professional Scoring

The open-ended reading and writing items were scored according to the procedures described in Chapter 8.2. For their initial scoring, the booklets were processed in the same order and session organization as they were received from data entry systems. However, scoring procedures required a reliability or second scoring for a 20 percent sample of the booklets. Accordingly, every fifth booklet in a batch was put aside for this purpose during initial scoring. Additionally, those booklets containing items to be holistically scored were held for that process while the remainder were forwarded to ETS key entry systems.

The back cover of each student booklet contained a row of boxes for each open-ended reading and writing item contained in the booklet. The boxes were used by scoring staff to enter scores and scorer identification codes according to the scoring specifications for each item. The primary trait scorers entered their identification codes into special boxes on the front cover of each booklet. Key entry staff transcribed the booklet serial number and scorer identification codes from the front of each booklet and the scores from the back. Because the number of boxes varied from item to item and the arrangement and number of items varied by booklet, the score data were loosely formatted on the data records, which were later untangled under the editing process. This untangling process is described in Section 8.4.2.

By the time the booklets had completed scoring and key entry processing, their session organization had been substantially corrupted. In anticipation of future writing assessments which would require re-scoring the writing items from Year 15, the booklets were reorganized and boxed by booklet number. This would facilitate the extraction of specific booklets from the raw data. The booklets were then shipped to the ETS data retention area for long-term storage.

8.1.5 Questionnaires

The questionnaire instruments were separated by type and accumulated by the operations coordinator as they were received from mail processing. These data were also transcribed through the data entry system but on a lower priority basis than the student booklets. The Excluded Student Questionnaires received higher priority than the Teacher and School Questionnaires, since the demographics of the excluded students were used in deriving the sampling weights of the assessed students. Every effort was made to keep the processing rate of these instruments in pace with the student data entry, in order to have the two files completed at the same time.

The Excluded Student, Teacher, and School Questionnaires each had their own processing options on the primary menu of the entry system. The entry operator would enter the appropriate code for an instrument and the entry mode to initiate processing. The questionnaire entry programs followed the same model as the student entry program with the absence of a tracking file and session batching. Entry, verification, and resolution modes were available; audit reports were initiated by the operations coordinator.

The Excluded Student Questionnaire entry program first displayed a screen for entry of the front cover data. The operator was prompted for the serial number of the booklet to be processed. An error condition occurred if either a record with that serial number was found under entry mode or no record was found under verification or resolution mode. In either case the operator was asked to verify that the correct number had been entered. If the problem persisted, it was referred to the operations coordinator for resolution. The remaining cover information, including PSU and school code, student sex, ethnicity, grade, and birth date, were processed as for the student booklet covers. The program then displayed a single screen for processing the responses within the questionnaire. When the operator pressed ENTER to terminate processing for that booklet, the program re-displayed the cover entry screen, ready to process another booklet. A blank field entered in the serial number field returned the program to the primary menu.

The Teacher Questionnaire entry program first displayed a screen for entry of the cover information. It processed the serial number in the same fashion as did the Excluded Student Questionnaire entry program. The cover information only included the PSU, school, and teacher codes. As the longest questionnaire instrument, the Teacher Questionnaire required three screens for entry processing due to software limitations as well as general appearance and ease of reading. Completion of processing for each booklet returned the program to the cover entry screen, where the entry of a blank serial number returned the program to the primary menu.

The School Questionnaire entry program also started with a display of the cover entry screen. The only information requested for this instrument, however, was the PSU and school code which also served as the booklet identification number. Entry processing for the questionnaire information was broken across two screens. Completion of processing for each booklet returned the program to the cover entry screen, where the entry of a blank PSU and school code returned the program to the primary menu.

After all questionnaires had been received and processed through the entry system, a final validation was performed on all data values in all records. Any data errors or discrepancies were corrected at this time using the resolution mode of the entry system. A final audit listing was generated, recording all entry activities for each questionnaire.

The questionnaires were subjected to the same quality control procedures that the student data received. The details of the sampling rates and results are discussed in Sections 8.5.2 through 8.5.4.

At the completion of quality control processing, the questionnaires were packed into boxes and shipped to the ETS data retention area for long-term storage.

Chapter 8.2

PROFESSIONAL SCORING

Anne Campbell

Educational Testing Service

The professional scoring of the Year 15 NAEP assessment was conducted for open-ended reading and writing items from all three grade/ages. Three methods of scoring were used: primary trait scoring for both writing and reading items, and holistic and mechanics scoring for writing items.

Although NAEP now scores writing responses mainly using the primary trait system, NAEP used holistic scoring for its first writing assessment in 1969. Holistic scoring evaluates responses on the basis of overall impression rather than on particular aspects such as mechanics or organization. As a relative process dependent upon the quality of writing received, holistic scoring did not completely address the need to report performance levels for particular writing skills or the need for a scoring system that could be replicated. As a result, NAEP began to search for an alternative scoring process to use in the next writing assessment.

With input from educators and measurement specialists, NAEP devised a system known as primary trait scoring. This system was designed to evaluate the ability to write for precisely defined purposes and thus uses closely defined tasks. When a writing item is developed, a dominant characteristic or primary trait is identified. This primary trait is the basis for establishing criteria for evaluating the responses. These criteria are associated with specific score points in a scoring guide. Each score point defines a level of task accomplishment, that is, the degree to which a response contains the characteristics required to accomplish the purpose of the writing task.

Although the primary trait system was developed specifically to evaluate responses to writing tasks, the scoring approach was adapted to evaluate responses to open-ending reading items. Criteria were defined to evaluate how well students responded to a reading passage when asked to perform such tasks as evaluating a story or poem, identifying and supporting a mood of a passage or using information in a passage to draw comparisons and contrasts. Criteria for each task were associated with specific score points in a scoring guide.

Two distinctions may be made about the items which were scored. First, all of the open-ended items were incorporated into booklets on the basis of the spiralling design. However, a few items were also used in booklets which were accompanied by paced audio tapes (see Chapter 5 for a discussion of spiral and tape administration.) In scoring the spiral and the tape booklets, no distinction was made between the two; the tape booklets were included with the spiral booklets in the batching process.

Second, to provide for trend analysis, four writing items were included in the Year 15 assessment which had been administered in previous assessments. Three of these items were from the Year 10 (1978-79) assessment; the fourth item had been used in both the Year 10 and the Year 5 (1973-74) assessments. Responses to these four items from the previous assessment years were not scored at the time they were collected, but were retained so that they could be scored at the same time and by the same scorers as the responses from the Year 15 assessment. Thus, when the scoring for Year 15 began, the responses from the previous assessment were intermingled and scored with those from the Year 15 assessment. These items were scored using both primary trait and holistic methods; a subsample of one of the items was scored for mechanics (described below).

Four reading items which had been administered in two previous assessments were also included in the Year 15 assessment to provide for trend analysis. The scoring of these items was handled in a different manner than the writing trend items. Responses to these items from the two previous assessment years were scored at the time they were collected and were then retained. When it came time to score the Year 15 responses to the same items, training papers from the previous assessments were provided for the readers to familiarize them with how the items were previously scored. Then a 20 percent subsample of the responses from the previous years was pulled, their scores were masked, and the responses were distributed to the readers who re-scored them. The previous scores were then unmasked and compared with the scores given by the current reader. If the scores of the Year 15 readers deviated drastically from the scores given previously, special training sessions were held to bring the readers into conformity with the previous scoring. During the time that this re-scoring was going on, the Year 15 responses to these items were also being scored.

The Year 15 NAEP assessment included the 35 open-ended items listed in Table 8.2(1). This table provides an overview of reading and writing items, including item number, grade/age level, and primary trait score ranges.

The rest of this chapter will describe the different methods of scoring and will discuss the scoring operation, including training, work flow, and reliability.

8.2.1 Description of the Scoring

8.2.1.1 Primary Trait Scoring

All open-ended reading and writing tasks were scored using the primary trait system of scoring. This involved assigning a score point based on a scoring guide designed for each item. The typical guide included score points of 0 to 4, 7, 8, and 9, although a few had score points of 0 to 3, 0 to 5, or 0 to 6, plus 7, 8, and 9. A general explanation of these score points is given below.

0, 7, 8, 9: These scores were given to responses that were blank, indecipherable, off task, or contained a statement to the effect that the student did not know how to do the task.

1: This score indicated an unsatisfactory response in that it was very abbreviated, circular, or disjointed and did not represent a basic attempt toward addressing the writing task.

2: This score was given to responses in which some or all the elements needed to complete the task were present but were not managed well enough to ensure that the purpose of the task would be achieved.

3: The responses given this score point included the information and ideas critical to accomplishing the underlying task and were considered likely to achieve the desired purpose.

4: This score was given to responses that went beyond the essential by providing more detail and being more coherent.

Along with scoring for the primary trait, some tasks also required the scoring of anywhere from one to four secondary traits (see Table 8.2(1) for tasks scored for secondary traits). The scoring of the secondary traits involved indicating the presence or absence of elements that were of special significance to that particular item. For writing items these secondary traits included whether or not notes were made before writing and whether or not critical information was filled in on a form. For the reading items, scoring for the secondary trait involved analyzing whether supporting evidence was based on content, form, or subjective reaction plus for some items indicating the number of pieces of evidence that were included. Primary and secondary trait scores for all items in a booklet were placed in designated boxes on the booklet's back cover.

Table 8.2(1)
Distribution of Reading and Writing Exercises

Item Name	NAEP Item Number	Reading Writing (R)/(W)	Grade/Age Use			Primary Trait Score Ranges	Secondary Traits	Holistic Score Ranges
			4/9	8/13	11/17			
Dali	N000100	W	X	X	X	0-9		0-6
School Rule	N000200	W	X	X	X	0-4,7,8,9		
Recreation Opportunities	N000300	W		X	X	0-4,7,8,9	1	
Food on The Frontier	N000400	W	X	X	X	0-4,7,8,9		
Dissecting Frogs	N000500	W		X		0-4,7,8,9		
XYZ	N000600	W	X	X		0-3,7,8,9		
Swimming Pool	N000700	W	X	X	X	0-4,7,8,9		
Pet	N000800	W	X	X		0-4,7,8,9		
Radio Station	N000900	W	X	X		0-4,7,8,9		
Appleby House	N001000	W	X	X	X	0-4,7,8,9		
Nuts	N001500	R	X	X	X	0-9	3	
Travels With Charley II	N001900	R		X	X	0-5,7,8,9	4	
The Door	N002300	R		X	X	0-9	9*	
Bethune	N002800	R	X	X	X	0-5,7,8,9		
Goods to Market	N003100	R	X	X	X	0-5,7,8,9		
Dependency	N003700	R	X	X	X	0-4,7,8,9		
Track Meet /Javelin	N004300	R		X	X	0-4,7,8,9		
Start to Work	N004600	R		X	X	0-5,7,8,9		
Hole In The Box	N007200	W	X	X	X	0-4,7,8,9		0-6
Childhood Memory	N007400	R		X	X	0-5,7,8,9	4	
Travels With Charley I	N007500	R		X	X	0-5,7,8,9	4	
Flashlight	N007600	W	X	X	X	0-4,7,8,9		
Ghost Story	N007700	W	X	X	X	0-4,7,8,9		
Favorite Music	N007900	W	X	X	X	0-4,7,8,9		
Split Session	N008000	W		X	X	0-4,7,8,9		0-6
Cow-Tail Switch	N008200	R		X	X	0-9	3	
Mother and Do	N008900	R	X			0-9	3	
Plants	N014700	W	X			0-3,7,8,9		
Spaceship	N014800	W	X			0-4,7,8,9		
Aunt May	N014900	W	X			0-4,7,8,9	1	0-6
High Tech Pizza	N015900	R			X	0-4,7,8,9		
Job Application	N019000	W			X	0-4,7,8,9	2	
Funding Space Center	N018000	W			X	0-4,7,8,9		
Uncle	N020000	W			X	0-4,7,8,9	1	
Bike Lane	N021000	W			X	0-4,7,8,9		

* The additional scores for "The Door" are not regarded as secondary traits, but as descriptive information about the student's response.

8.2.1.2 Holistic Scoring

Four items were also scored holistically. These were items planned for use in the trend analysis and so included responses from Years 5 and 10 as well as from Year 15. The responses for each task for each age were randomly mixed together and rated relative to each other. The holistic scoring was performed as a separate task from the primary trait scoring by a different group of scorers. Holistic scorers evaluated each response according to overall impression, then assigned scores from 1 to 6 (with a special score for papers that were blank or unrateable). Holistic scores were placed in designated boxes on the back covers of the booklets.

8.2.1.3 Mechanics Scoring

In addition to primary trait and holistic scoring, a third procedure, scoring for mechanics, was applied to a subsample of responses from the exercise "Hole in the Box." Five hundred essays were selected from each age for each of the three assessment years in which the exercise was administered. Each group of 500 essays selected for each age included responses from 200 students who were black and 300 students who were not.

The responses were duplicated with the student identification number indicated on the copy. They were bundled by age in such a manner that responses from the three assessment years were randomly mixed. The mechanics scoring evaluated the elements of sentence construction, word choice, spelling, punctuation, and capitalization. To do this, a reader wrote symbols in red ink at each word or punctuation mark in error and at the ends of sentences to indicate sentence type or faulty sentence construction.

To analyze the data from the mechanics scoring, criteria were devised to derive scores from mechanical scoring codes. The codes included:

(1) the number of words in an essay;
(2) the number of sentences in an essay;
(3) the number of letters in a word;
(4) the number of "T-Units";
(5) sentence construction; and
(6) punctuation.

These criteria are described below.

Number of Words in an Essay. Each blank space used in key entry of an essay counted as one word. For errors that occurred when a student separated one word into two (e.g., "mail man" for "mailman"), readers enclosed the error in brackets, to indicate that the two words should be counted as one.

Words which could not be deciphered were circled by readers and followed by the letter "L." The "L" was keypunched with the essay to indicate that the circled material should be counted as one word.

Number of Sentences in an Essay. Certain mechanical scoring codes were used at the end of a sentence. After keypunching was completed for an essay, these codes were tallied; the total counted as the number of sentences in an essay.

Number of Letters in a Word. The mean length of the words used by a student was determined by dividing the number of letters used to keypunch an essay by the number of words in an essay.

T-Units. NAEP uses T-Units to assess the quality of syntax used in an essay. A T-Unit is an independent clause and all of its modifying words, phrases, and clauses. T-Unit counts were calculated as follows:

(1) a simple sentence counted as 1 T-Unit;
(2) a complex sentence counted as 1 T-Unit;
(3) a compound sentence counted as 2 T-Units;
(4) a sentence fragment was added to a following sentence so that it became a clause (constituting a T-Unit) in the new sentence; and
(5) a run-on sentence constituted several T-Units, depending upon the number of clauses it contained.

Sentence Construction. To assess further the quality of syntax used by students, the following were calculated:

(1) percent of simple sentences;
(2) percent of compound sentences;
(3) percent of complex sentences;
(4) percent of sentence fragments; and
(5) percent of run-on sentences.

To determine the number of instances of faulty sentence construction, the following were calculated:

(1) the average number and percent of sentences with agreement errors (obtained by dividing the number of "A"s assigned to an essay by a reader by the number of sentences in that essay. The letter "A" is used to signify agreement errors in sentence construction.)

(2) the number of errors in word choice; and

(3) the number and percent of sentences that were considered awkward.

Punctuation. Counts were obtained for the following errors:

(1) the average number and percent of misspelled words; and
(2) the average number of errors in capitalization.

Punctuation errors were divided into three categories:

(1) errors involving commas and dashes;
(2) errors involving end marks (periods, question marks, and exclamation points); and
(3) errors involving other forms of punctuation.

and were calculated by:

(1) errors of commission for each of the three categories above and for overall punctuation; and
(2) errors of omission for each of the three categories above and for overall punctuation.

In addition to the data specified above, NAEP obtained a summary of all mechanical errors for "good" and "poor" essays. The terms "good" and "poor" refer to the primary trait scores assigned to each essay.

8.2.2 The Scoring Operation

8.2.2.1 Scorers

Fourteen persons were hired specifically to score the NAEP reading and writing exercises using primary trait scoring. The same fourteen persons also performed the mechanics scoring.

Generally, the persons chosen had teaching experience ranging from the pre-school to the community college level. The group included men and women of various ages and racial/ethnic groups who had lived and/or gone to school in various parts of the country, and who had BA and MA degrees (a few were working toward doctoral degrees). The persons who performed the holistic scoring were required to be presently teaching.

8.2.2.2 Training: Primary Trait Scoring

Before the training of the scorers began, NAEP staff worked with the scoring coordinator and assistant coordinator to prepare training sets and to refine the scoring guides.

Training began with the 26 items administered to Grade 8/Age 13. This training involved explaining the item and its scoring guide, discussing responses that were representative of the various score points in the guide, then scoring and discussing approximately 65 to 100 randomly selected responses. The purpose of the training was to familiarize the group with the scoring guides and to reach a high level of agreement among the scorers. After the group training was completed, each scorer scored the items in each of fourteen bundles of booklets. Their scores were recorded and a follow-up session was held to discuss those responses for which there was a wide range of scores. Once the follow-up session was completed, the scoring began. Initial training was completed in approximately one month.

As a follow-up to training, notes on various items were compiled and distributed to the scorers for their reference. In addition, short training sessions were conducted on items that showed low reliability. The scoring supervisor consulted with individual scorers as the scoring progressed. When a scorer was judged to be causing a discrepancy, the supervisor would discuss the response and its score with that scorer.

As scoring began for each of the other two grade/age levels, training was conducted on the items unique to those levels. The training was the same as that conducted initially and took about one week for each grade/age level.

8.2.2.3 Training: Holistic Scoring

The training for holistic scoring involved several steps. First, the table leaders--all of whom were experienced holistic readers--surveyed the pool of papers from assessments and selected anchor papers, that is, papers representative of six levels of proficiency. Then, they developed guidelines describing each level and how to distinguish between top-half and bottom-half papers. The training began with some discussion of the characteristics of the anchor papers and guidelines, then included several practice scorings of other papers to refine the scoring scale description and to resolve discrepancies among readers. When all readers were comfortable with the guidelines, they scored papers for an hour, after which they discussed additional anchor papers. Throughout the subsequent scoring there were periodic discussions of papers to ensure that readers continued to adhere to the same standards.

8.2.2.4 Training: Mechanics Scoring

To prepare for mechanics training, the scoring coordinator and an outside consultant with experience in mechanics scoring refined the guidelines and selected papers to be used in training. The training itself involved discussing the guidelines and sample responses which had already been scored. The scorers then practiced scoring other papers, and discussion was held when any discrepancies occurred. When readers were comfortable with the guidelines, the actual scoring began. Several follow-up training sessions were conducted as problems arose.

8.2.2.5 Assignment of Work

For the primary trait scoring, the scorers received the booklets in batches as they were received from the schools. A reader scored all open-ended items in all booklets of a batch. Because of the spiral design, a reader would encounter many, if not all, of the items at a grade/age level as he or she scored a batch of booklets. Thus the reader had continual exposure to all items throughout the scoring. Interspersed among the batches of Year 15 booklets were the responses for several items from the two other assessment years. The responses for each item were bundled together in groups of 25 by age and by assessment year.

The three grade/age levels were scored separately, beginning with Grade 8/Age 13, continuing with Grade 4/Age 9, and ending with Grade 11/Age 17. It was hypothesized that this procedure may have led to a "batch effect": that is, the Grade 4/Age 9 essays may have been evaluated as too high because, after reading the essays written for Grade 8/Age 13, scorers may have considered the Grade 4/Age 9 responses "pretty good for fourth graders." Correspondingly, Grade 11/Age 17 responses may have been rated too low because, following the Grade 4/Age 9 responses, they may not have seemed "that good for eleventh graders."

To determine the effect of scoring the papers in batches by grade/age levels, an experiment was performed in which NAEP written responses from all three grade/age levels were randomly ordered, then re-scored. It was decided that if batch effects exceeded one-tenth of a score point per item, post hoc adjustments of the writing scale values would be warranted.

The experiment was based on responses to three writing tasks that were administered to all three grade levels--School Rule, Food on the Frontier, and Swimming Pool. For each writing task at each grade/age level, a [illegible]presentative subsample of 156 to 174 papers was drawn. Because the booklets administered to each grade/age level were different colors, the responses were photocopied, then reordered using a randomly selected permutation of their sequence numbers. The responses were then scored by two experienced readers. The data were analyzed, and it was concluded that no adjustment of writing scale values was required. (See Chapter 11.1 for more information on the batching effect.)

The other two scoring procedures were performed under basically the same conditions. Each cohort was scored separately and responses from the two previous assessment years were mixed in with those from Year 15. However, the holistic scorers received only those booklets which had the holistic items in them and so did not receive the booklets in batches as they came from the schools. For the mechanics scoring, the readers received photocopies of the responses.

8.2.2.6 Reliability and Resolution

Twenty percent of the primary trait items were subject to a reliability check, which entailed a second reading by a different scorer. To prevent a second reader from being influenced by the first reader's scores, the first reader masked all the scores in every fifth booklet in a batch. These booklets were passed along to a second reader, who scored for the primary trait only. All scoring discrepancies were independently resolved by the scoring supervisors who assigned a resolution score. In most instances, this score was the same as one of the given scores. However, in a few cases, neither score was considered correct and so a different score was given. Although the secondary trait scores were not subject to a reliability check, they were sometimes adjusted by the scoring supervisor to maintain consistency with the resolved primary trait score. (See Chapter 11.1 for a description of the results of the reliability check.)

Holistically scored items were also subject to a 20 percent reliability check. The scores of the first reader were masked and the papers were passed on to a second reader. When discrepancies occurred, alternating high and low scores were assigned if the scores were one point apart. That is, if the first occasion of a discrepancy of one point was resolved by assigning the lower of the two scores given to an essay, the next occasion of a discrepancy of one point was resolved by assigning the higher score. Discrepancies of two or more points were resolved by the scoring director, as in the case of the primary trait scores.

The same general procedures were followed for the mechanics scoring: 20 percent of the responses were re-scored; second scorers did not see the first scores; and discrepancies were resolved by the scoring supervisor.

8.2.2.7 Data Entry

After the scoring was completed, the booklets were sent to data entry, where the scorer ID numbers from the front cover and the scores from the back covers were entered. (See Chapter 8.3 for details concerning the data entry process and Chapter 8.4 for information concerning editing data.) The booklets went to key entry in batches except for booklets which had items for holistic scoring; these were pulled from the batches and held. After holistic scoring was completed, those booklets were sent to key entry.

Chapter 8.3

DATA ENTRY SYSTEM

Alfred M. Rogers

Educational Testing Service

The transcription of response data from paper to machine-readable form is one of the most important yet often overlooked aspects of any research project. Among the many issues to be considered are the collection, delivery, and management of the physical data; the actual machinery, including hardware and software, employed for the transcription process; the validation of the machine-resident data; and the management of the data files.

In terms of volume of data collected, the Year 15 NAEP was comparable to that of most administrations of the large testing programs at ETS and within the capacity of extant data transcription technology. However, the BIB design and the spiralled distribution of its many booklets created a complexity beyond the capability of that technology. A new methodology was developed for the sole purpose of transcribing the NAEP data into computer-readable form. This chapter traces the development and implementation of this system from a discussion of its requirements, through a description of its design, to a detailed exposition of its operation.

Figure 8.3-1 is a schematic diagram representing the processing flow of student assessment materials through the data entry system. The reader may refer to this diagram for clarification of the relationships among the components of this system.

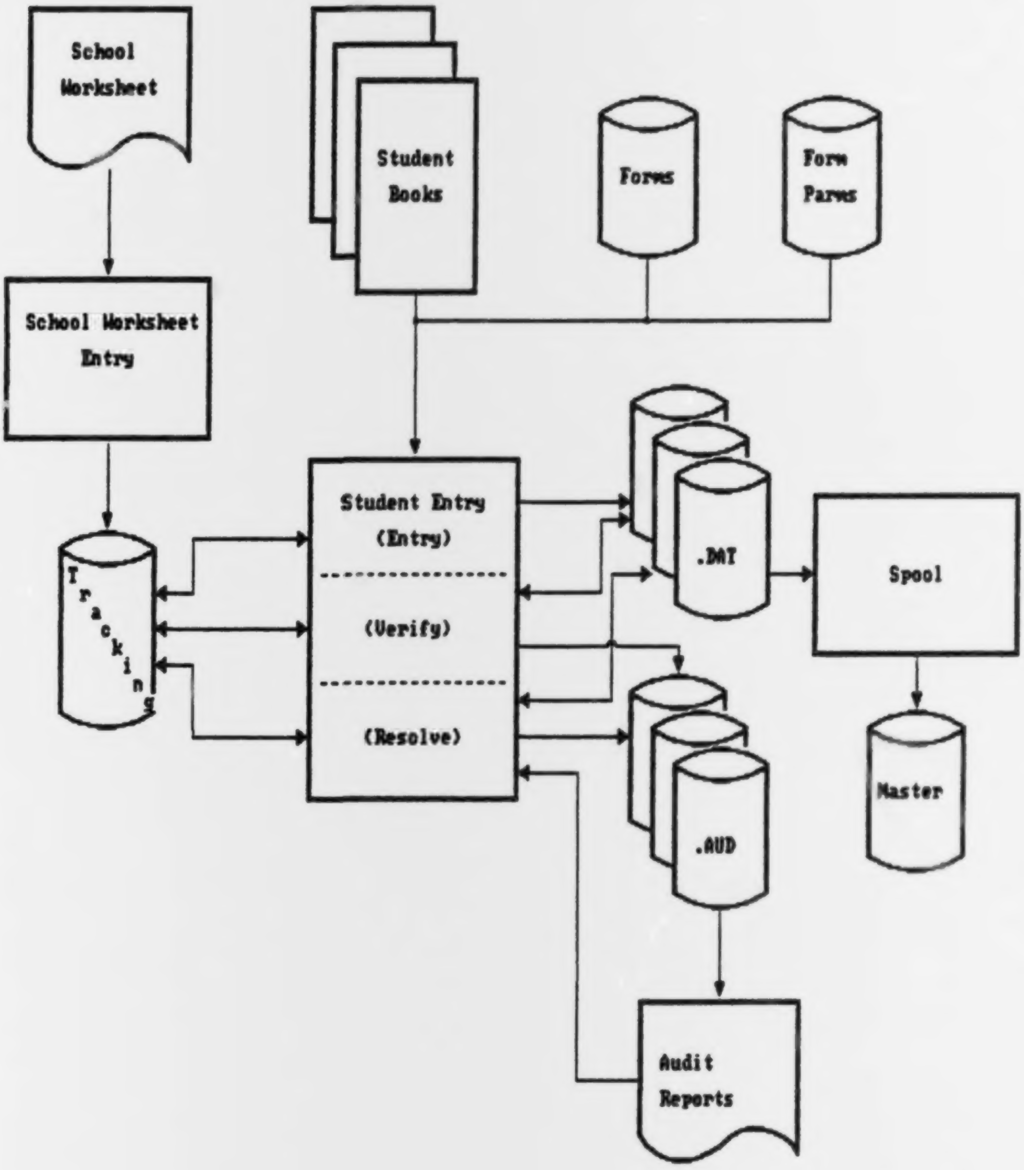
8.3.1 System Requirements

The primary consideration in the design of any data transcription scheme should be the interaction between the entry operator and the machine. An effective system should provide direct access to the data, a convenient entry and editing mechanism, and an accurate display or representation of the data values. At the next level, the system should provide data and file management capabilities and error detection and correction functions. Finally, a complete system should also include status reporting and quality control procedures.

The data terminal provides two interface components between the operator and the machine: the keyboard and the display device. The arrangement and function of the keys on the keyboard are critical to the

Figure 8.3-1

Student Data Entry Processing


School Worksheet
Student Books
Forms
Form Params
School Worksheet Entry
Student Entry
(Entry)
(Verify)
(Resolve)
Tracking
.DAT
Spool
Master
.AUD
Audit Reports

speed and accuracy with which data and commands can be entered by an operator. A numeric keypad is preferable to the typewriter keyboard because it allows the operator to use one stationary hand for entry while freeing the other hand for page turning or other tasks. If, as in the case of the NAEP instruments, the responses are coded in alphabetic format rather than numeric, the keypad numbers may be converted to letters through program control, rendering the keypad a more powerful entry device. Unless the entry operator is a skilled typist, the entry of alphabetic and special characters can slow entry processing considerably by adding key search time and intermittent hand movement from keyboard to booklet.

The manner in which data are displayed can also have significant impact on the efficiency of the entry process. The most primitive mode of display is the line editor mode, in which data appears as a continuous string of characters, wrapping around as many lines of the display device as are required to display the entire record. This puts a considerable burden on the entry operator to be able to identify a value in the data record from its location in the instrument. Even if the displayed lines were enhanced with rulers indicating column positions, the operator would be required to know the correspondence between column position and item number.

A more desirable alternative is the full screen mode, which uses panels or forms as the data input and display mechanism. A form may be regarded as a template consisting of protected areas (text) and unprotected areas (fields). The unprotected areas of a form are the "holes" in the template where data values may be written into or read by an application program. The protected area is the "body" of the template which cannot be accessed by the application. A form designed for the entry and display of data could have a separate field for each data value, a description of each field in an adjacent text area, and both text and field arranged in a logical order consistent with the layout of the instrument.

Since human eye-hand coordination is subject to error, any data entry system operated by humans should provide three modes of operation: entry, verification, and resolution. The entry mode takes the operator's keystrokes, validates them for data type and value range, and creates the data record. The verification mode takes the operator's keystrokes, validates them again, and compares them with the data values on the previously written record. It should notify the operator of any inconsistency and permit over-writing of the field if the initial value is determined to be in error. The resolution mode displays the current data values and permits the selection and correction of any field.

These basic requirements are complicated by the special demands of the NAEP design. The spiral design combines a relatively small number of subtests or blocks into a multitude of booklets. any booklet-oriented entry system would need one data format for each booklet and require that all booklets of the same type be batched together for efficient processing. the spiralled distribution of the assessment booklets renders this approach impracticable, if not impossible, since the incoming bundles of booklets would first have to be separated into piles of like booklets and entry processing would have to be delayed until there were enough booklets in a pile to make the effort productive. the separation process would also disrupt the session identity of the booklets, and special care would be required to insure that their proper identification.

A more appropriate entry system would maintain this session identity by processing the incoming bundles as complete units. As each booklet within the batch were presented for processing, the system would determine the format to be used according to the booklet identification code. Since there are fewer block types than booklet types, a more logical, and economical, extension of this concept would be to treat the booklet format as a sequence of block formats.

8.3.2 Machine Considerations

Due to the time and budget constraints between the awarding of the naep grant and the field administrations, it was not possible to develop scannable or machine-readable booklets or answer sheets. The assessment instruments were marked or written in and manually transcribed to machine-readable form. Conventional key entry systems, which process at the booklet level, were ruled out for the reasons mentioned above. At the time of preparation for data collection, ets was installing a more sophisticated key entry system which could be programmed to operate at the block format level. However, it was not anticipated that the expertise in using this system could be developed before entry processing would begin, nor was it known whether the system could handle the number of block formats.

The remaining alternative was a computer-based entry system, or more appropriately, an interactive program for data generation and management operating on a mini- or mainframe computer system. The only computers available for use were a VAX 11-780 system running under VMS, and an IBM 3083 system running under OS/MVS-TSO.

Both machines offer very similar programming environments for the development and implementation of interactive systems: full-screen editors for program code and control data creation and modification; assemblers, FORTRAN compilers, and linkage editors for program construction; forms management systems with editors, utilities, and callable interfaces with other languages; and direct-access data storage with sequential, library, and indexed data structures. The data terminals for both machines provide similar environments for the entry operator: a full typewriter-style keyboard, numeric keypad, and function keys.

The IBM Time Sharing Option (TSO) is a multi-user interactive subsystem with full capabilities for program creation, testing, and implementation. The System Productivity Facility (SPF) subset of TSO is a menu-driven, full-screen utility for the creation, editing, and maintenance of program code and data files. The Dialog Manager Service (DMS) allows the program developer to design and use SPF-like panels in full-screen application programs. The FORTRAN H compiler conforms to the 1966 standards, with no capability for processing CHARACTER-type data or dynamic file allocation as in the 1977 standards. Neither does it provide any capability for processing library files or indexed files, both of which are valid data structures on the IBM system.

The VAX VMS operating system is designed as an interactive user environment with much the same capabilities as TSO. The EDT editor may be used for editing both program source and data files. The Forms Management System (FMS) is a separate product which provides its own forms editor, library management utilities, and callable interfaces for full-screen program development. The FORTRAN compiler conforms to the 1977 standards, with CHARACTER-type data and dynamic file allocation, and interfaces with library and indexed files. Additionally, two other products available on the VAX had great potential for higher order management functions: the Common Data Dictionary (CDD) which could store information about data files and record structures, and DATATRIEVE for the retrieval and reporting of information within the CDD.

From a management standpoint, the IBM was preferable to the VAX, not only because the NAEP data analysis would be performed on that machine, but because the rates were more favorable. One important factor to be considered was utilization. The IBM TSO was heavily used during prime shift hours (8:00 a.m. to 4:00 p.m.) and experienced performance slowdowns during peak activity hours. The VAX was implemented as a research and development tool and had no production load to contend with. Although it was this difference in utilization which accounted for the discrepancy in rates, it made the VAX more attractive for its stability.

From a programming standpoint, the VAX offered the most functionality and flexibility in developing an interactive data entry system. The lack of file management interfaces under the IBM FORTRAN constrained the program developer in choice of file structures or forced the development of new interfaces or structures.

Ultimately, however, it was from an operational standpoint that the VAX was chosen. The forms input and output functions on the IBM worked at the screen level; that is, the calling program would issue a read command to the terminal and wait until the operator had pressed a function key, at which point the contents of all data fields would be returned. The forms input and output functions on the VAX worked at the field level; the contents of each data field on the form could be accessed and processed individually, before passing control to the next data field. This meant that data could be captured at the keystroke level, validated by the program, and either continue to the next field or notify the operator of a problem. The IBM program could only perform the validation after all fields had been entered, forcing the operator to go back through the booklet for any subsequent error processing.

This field-level access mode also made it possible to process fields on a conditional basis. Several background and questionnaire items in the assessment had a "Specify Other" option in which the respondent circled the letter preceding the option and filled in a short written response. With special coding the program could be instructed to capture data from the open response field if the option was selected or to bypass the field if another or no option was chosen.

The VAX terminals also offered a "true" numeric keypad as opposed to the "function" keypads on the IBM. It would have been possible to use certain keys on the typewriter keyboard to emulate numeric input if the need arose, but that, too, would have placed additional demands on the entry operators.

8.3.3 Database Organization and Structure

The organization and internal structure of the data and control files is the framework around which the entry system is built. A knowledge of the structure and purpose of these files, as well as their relationships to each other, is central to an understanding of the entry system.

The storage of the transcribed data is always the first consideration. An indexed data file using a unique identification code such as the booklet number as the access key is the most accurate and direct means of storing and retrieving data records. A single, albeit large, indexed data file is conceptually the easiest solution to the needs of data storage and access, but creates other problems from a management perspective. The records within an indexed data file are stored in ascending key order. As records are added to the file, they are inserted between existing records to preserve the sort order, and the pointers to these records are updated. Under the spiral design, any spiral session contains a wide assortment of booklet numbers. Inserting these records into an indexed master file would not only incur additional overhead processing for reorganizing the data file, but disrupt, if not destroy, the session identity of the booklets.

An alternative solution would be to store the data in smaller "batch" files, borrowing a term from optical scanning and key entry methodology. The session is a logical choice for a unit of batch processing: all field administration management functions were done at the session level; the session sizes are fairly consistent at about 23 to 27 booklets each; and each session is uniquely identified by its PSU, school and session codes. In any case, maintaining session identity of the booklets was a primary consideration. The data records within a batch file would still be stored in key order, but once they were written to the file, any subsequent activity would not require reorganization.

Having hundreds of batch files on the computer necessitated some means of keeping track of each batch as it went through the entry system. A tracking file was designed in which each record would store the processing history of a single batch data file. The tracking file is also an indexed file, using the session identification code as the access key. This one-to-one correspondence between tracking file record and batch data file is central to the processing and management functions of the student data entry system.

Because the verification and resolution processes are capable of altering data values in the batch files, an audit trail mechanism is required to trace the evolution of the data through these procedures. For that purpose, a separate audit file is maintained for each batch data file on the system. Any time an anomalous data value is detected, or a data value is altered by either the verification or resolution process, a record is written to the audit file, giving complete information about the booklet, section, and item number for the field, the old and new data values, and the date and time the action occurred. These files are organized as sequential files, to which each audit record is appended during the different stages of the entry process.

The data entry system required these three types of files for data storage and processing control. The actual operation of the entry program required two additional files: the forms library and the forms parameters file.

The forms library stores all of the forms used in the entry system. The forms are created and updated using the FMS editor. The library is updated and maintained using FMS utilities. The forms are accessed by the entry programs through the FMS forms driver routines.

The forms parameters file is designed as an adjunct to the forms library for the control of field processing within each form. Each record in the parameters file corresponds to one field in one form. The parameters file is organized as an indexed file using the form name and field sequence number as the access key. The integration of the forms library and the form parameters file will be elaborated later.

8.3.4 Program Structure and Execution

The data entry system, as used by the entry operator, is a single FORTRAN-written program with special subprograms to handle the various components. The program is initiated by a single command from the entry operator.

The program's first task is to define its operating environment for recording on the tracking file and audit trails. The date, time, and terminal address can be obtained using system-resident functions, but the operator identification must be requested from the operator. When the operator code is entered, the program displays the primary options menu form. This form contains two fields to be filled in by the operator and a listing of the options and their numeric codes below each field. The first field to be filled is the OPTION code, indicating which instrument is to be processed: the school worksheet, student data, Excluded Student Questionnaire, Teacher Questionnaire, or School Questionnaire. The second field is the MODE code, indicating whether the selected instrument is to be processed for entry, verification, or resolution.

The definition of the program environment continues with the validation and storage of the entered codes. The environmental and other control parameters are stored in a COMMON data area for use by the other components of the system. The forms library, form parameters file, and tracking file are then opened or readied for processing. The program then transfers control to one of the five subprograms corresponding to the OPTION selected.

8.3.5 School Worksheet Processing

The school worksheet entry subprogram performs two functions: it provides for the entry of session administration information from the school worksheet, and initiates processing of the data for that session. This program uses two forms for the collection of data. The first form requests the school identification code and the number of spiral and tape sessions administered in that school. The total number of sessions determines the number of times the second form is used for the entry of session-specific information. The data from each column of the school worksheet is entered and stored on a separate record on the tracking file.

As mentioned above, the tracking file is indexed, using the school and session code as the access key. To insure that this key is unique, the sessions within each school are assigned codes accordingly: regular spiral sessions are assigned codes from 01 to 10; regular tape sessions are assigned the booklet number used in the session, in the range of 64 to 67; makeup spiral sessions are assigned codes 11 to 15; and makeup tape sessions receive the value of the booklet number plus ten, in the range of 74 to 77.

The remainder of the tracking record is initialized to blanks for the date and time stamp fields, and zeros for the count fields. The record is then written to the tracking file, ready for the entry of student data for that session.

The school worksheet program is the only one of the five to operate in entry-only mode. The verification of the worksheet information was not as critical as the possibility that subsequent processing of the tracking record might contaminate the control field information. For this reason, the operations coordinator was given the capability to alter tracking file information.

8.3.6 Student Data Processing

The student data entry sub-program is initiated by selecting option number two on the primary menu. The first form requests input of the identification code of the session to be processed. The program issues a read to the tracking file for the record corresponding to that session. If the record is not present, either the school worksheet information has not been entered for the session, or the operator has incorrectly entered the session code. A warning message prompts the operator to correct the code, enter another session code, or return to the primary menu.

If the tracking record is found, the program reads the control fields to determine the last activity performed on that session, and compares it with the processing mode specified in the primary menu. If the current mode is equal to or greater than the last activity code, the operator is allowed to continue. For example, if the last activity performed was verification and the current mode is also verification, the program assumes that previous verification processing was interrupted and is to be resumed. If, on the other hand, the current mode is entry, the program insists that entry has been completed with the initiation of verification. In this situation, the operator may not process this batch and must either return to the primary menu to change modes, or select another session to process under the current mode.

At this point, if the current mode is either verification or entry, the program reads the vector of booklet counts from the appropriate control area in the tracking record. These counts will be updated by subsequent processing and rewritten to the tracking record at the completion of processing. The batch data and audit files are then opened for input and output processing. The booklet cover form is displayed, requesting input of the student ID code for the booklet to be processed.

Upon entry of the six-digit code, the program issues a read to the data file for the data record corresponding to that booklet. An error message is issued if either: the data record is found and the current mode is entry; or the record is not found and the mode is verification or resolution. In either case, the operator may correct the booklet code, enter a new booklet code, or return to the session entry form.

To ensure that the correct booklet number has been entered, the operator is prompted to enter the single-letter block codes printed on the booklet cover. The program will not proceed unless the correct block codes have been entered, since these codes correspond to the formats to be used in processing that booklet's data record. By definition, all of the booklet numbers in a tape session must correspond to the session code, therefore no block validation is performed for tape booklets.

If no record is found under the entry mode, the program sets up to create a new record. The operator is prompted to enter the remaining fields from the booklet cover: administration code, grade, exercise administrator code, sex, race, birth date, and school code.

If a record is found under the verification mode, the program sets up to accept input as if it were in entry mode. However, as each field value is entered, it is compared against its corresponding location on the data record. If the values agree, processing continues with the next field. If they disagree, a warning message is issued and the program "locks" on that field, giving the operator an opportunity to determine and enter the "correct" value. A more complete explanation of the verification process is given below.

If a record is found under the resolution mode, the program displays the front cover data values from the record in their corresponding fields in the form. The operator may then use the TAB and BACKSPACE keys to move from field to field and overwrite any field value.

The operator presses the ENTER key to terminate processing of the front cover form. If there are any blank or partially blank fields on the form, the program signals that entry is not complete and the operator must fill those fields with either valid data values or the missing data code. If the form is complete, the program prepares to process the sections within the booklet. The program uses a control table organized by booklet number and section number to determine which blocks correspond to each section in each booklet. For each section, control is passed to the FORM_ENTRY subroutine to complete processing of the data record.

8.3.7 Forms Processing

The FORM_ENTRY routine is the workhorse of the data entry system, and serves as the model for all other full-screen entry functions. It receives from the calling program the name of the block to be processed and a work area. In the entry mode, this work area is received as a string of blank characters and returned to the calling program as a contiguous string of entered data values. In the verification and resolution modes, it contains the data string from the input data record and returns the modified data to be written back to the data record. The routine also returns to the calling program the length of the data string.

The block name received by the routine is a two-character mnemonic code assigned to the cognitive and background item blocks. It also identifies which form to use to process the response data for that block. For the student data, all blocks contained few enough items to be represented in a single form without a cluttered or crowded appearance.

The items within the block are arranged in column order on the form, using three or four columns of approximately equal length. Each item is labeled in the text area by its sequence number within the block, followed by as many data fields as there are possible responses to that item. Each data field is named according to the NAEP number printed beside its corresponding item. This field name does not appear in the displayed form, but is used as an internal identification code by the forms management system. The data fields were "flagged" by an underline attribute to distinguish the data entry and display areas from the text part of the form.

An application program accesses a field within a displayed form only by using the field name. The application must therefore "know" the field names within a form, how they are to be processed, and in what order they are to be processed. This information is provided to the entry system by the forms parameter file.

The forms parameter file contains one record for each field for each form. The file is structured as an indexed file, using the form name and sequence number of the field within the block as the access key. After loading in the designated form from the form library, the routine locates and reads the record corresponding to the first field from the parameter file using an indexed read. The remaining records are read sequentially from that point until a record for another form or end-of-file is encountered. The parameters on these records are loaded into an internal table which is used by the routine in processing the data for this block. The contents of the parameter table will be listed here and their functions elaborated below: item number, field name, alternate form, alternate field, field type, field width, number of valid responses, next field name, conditional codes and conditional field names.

After the form and its parameters have been loaded, the routine determines its processing environment from the control parameters in the common area. It displays the current booklet section number as part of the form title and the processing mode in the lower right corner. Entry processing begins by setting an index to point to the first field on the form. Since all fields are processed in an identical manner, it suffices to describe the processing of a single field.

The program "reads" a field by invoking an FMS-supplied routine which uses the field name as input and returns the contents of the field and a field terminator code. There are four terminator codes recognized by the routine, three of which correspond to function keys on the keyboard: ENTER, TAB, and BACKSPACE. The fourth terminator code, AUTOTAB, indicates that the field has been completely filled by operator input.

The ENTER code indicates that the operator has pressed the RETURN or ENTER key to terminate processing for the form. The program scans the form for blank data fields to ensure that all fields have been processed under the entry and verification modes. If a blank is found, the program issues a warning message and the operator must complete the form to proceed with the next section.

The TAB and BACKSPACE codes indicate that the operator has pressed their corresponding keys to move ahead one field or back one field, respectively. The field pointer index is either incremented or decremented and the next or previous field is processed.

If the AUTOTAB code is returned, the entry operator has made one or more keystrokes to fill the requested field. The field width parameter corresponds to the size of the field in the form and indicates the number of characters returned for processing. The field type parameter indicates how the returned data is to be processed. The data fields on all forms fall into one of four types:

Type 1 - All of the multiple-choice, single-response items. The responses are coded by letters rather than numbers. All numeric input data values must be translated into their corresponding letter codes before being output to the form and data record. These fields are also subject to range validation.

Type 2 - All numeric data which may be checked for value range. This includes the "circle all that apply" items and numeric codes for some of the open-ended-response items.

Type 3 - Any numeric data which can only be validated for numeric type but not for range. This data includes counts and percentages.

Type 4 - All open-ended-response items which cannot be codified and must be represented in their raw form. These fields are always eight columns in length and may contain any combination of alphabetic, numeric, or blank characters.

The returned data value is first compared against three values designated by the three non-numeric codes on the keypad. The hyphen is used to indicate "no response" to the item. This code is valid for all field types. The period indicates that two or more choices were selected where only one choice was permitted. This code is only valid for the first two item types. The comma, translated into a question mark by the program, indicates a response which cannot be resolved by the entry operator and requires coordinator intervention. This code is valid for all field types.

If the data value does not meet the above criteria, it is then processed for validation. If it is one of the first two types, it is validated for range according to the number of valid responses parameter. If it is a Type 3 field, it is checked for numeric only. A Type 4 field has all blank characters translated into underline characters, because the end-of-form processing does not allow blanks in the returned data string. If the field contains an invalid data value, the program issues a warning message at the bottom of the screen and "locks" onto this field. The operator must enter a valid data value before the program will continue processing another field. Even the use of the function keys is prevented until a valid value is entered.

If the entry mode is indicated, the program writes the data value into the next available location in the work area and sets up to process the next field. If it is operating under verification, the program compares the entered data value with its corresponding location in the work area. If the values agree, processing resumes with the next field. If they disagree, the program issues a warning to the operator and again locks onto the field. The program will not release control of the field until the operator presses the ENTER key, indicating that the "correct" value has been entered. The program then writes the new value into the work area and continues with field processing. If in the resolution mode, the program over-writes the work area with the input value.

The "next-field" parameter contains the name of the next field to be processed after the current field. If the field has any non-blank conditional codes, the program first compares the entered data value with these codes. If there is a match, the corresponding conditional field parameter is used instead of the next-field parameter. The field pointer index is incremented and the corresponding field name is compared with the next field name. If they do not match, the field on the form and its corresponding work area location are filled with the "no response" code and the index incremented again until a match is found. The last field on a form is signaled to the program by a next-field parameter code of "LAST", at which point a message is issued to the operator indicating the end of the form.

The alternate panel and alternate-field parameters are used for the few open-ended items with codeable responses. An alternate form was generated for each of these items containing a listing of the possible responses and their corresponding codes. The alternate-field parameter indicates the field name on the alternate form to receive the data value. On completion of processing for this field, the original form is re-displayed with the new data value in its field.

8.3.8 Audit Trail Processing

The audit file for each session contains information on the processing history of selected data fields within the session data file. The entry programs and routines write a record to the audit file when the following field processing conditions have occurred:

(1) Under entry mode, the multiple response code was entered.
(2) Under any mode, the unresolvable code was entered.
(3) Under verification and resolution modes, a data value was written to the work area which differed from the original value.

Each record contains an identification section, consisting of the school and session codes, the booklet serial number, section number, and item code; a processing section consisting of the operational mode, the old data value, if applicable, and the new data value; and an environment section including the date, time, and operator code.

The session entry processing terminates with the production of an audit trail report. The program produces a formatted listing of the audit file contents at a printer located near the entry terminals. This permitted the operator to enclose the report with the session materials for later processing.

8.3.9 Questionnaire Processing

The Excluded Student, Teacher, and School Questionnaire data entry functions are performed by three separate sub-programs, each invoked by different option codes on the primary menu form. Since the processing of each instrument has more similarities than differences, the entry procedures for all three will be described in this section and differences will be noted where appropriate.

The data for each instrument are maintained on single, indexed files using the booklet identification code as the access key. The audit trail for each data file is also maintained on a single data file, which constrains entry operation to one operator at a time for each instrument. The booklet cover entry form is first displayed, requesting entry of the booklet number. The program issues a read to the data file using the booklet number as access key. An error message is issued if either the data record is found under the entry mode or the record is not found under verification or resolution modes. In either case, the operator must check and enter the correct identification or return to the primary menu to change modes.

Front cover processing continues with the entry of the remaining information. On the Excluded Student Questionnaire, the grade, sex, ethnic code, birth date, and PSU and school code must be entered. On the Teacher Questionnaire, only the PSU and school code and the teacher identification code are input. On the School Questionnaire, the PSU and school code serves as the booklet identification code so no other data fields are required. The operator presses the ENTER key to terminate front cover processing. The program invokes the FORM_ENTRY subroutine to process the response data as if the questionnaires were composed of separate sections. The Excluded Student Questionnaire has few enough data fields to be contained on one form, but the Teacher and School Questionnaires had to be broken across three and two forms, respectively.

Audit trail reporting is not activated at the conclusion of processing for each instrument. This function was provided to the operations coordinator to be performed on a periodic basis. The audit report program for each instrument would first sort the audit file by booklet and item number to facilitate the location of specific booklet numbers in the voluminous printed output.

8.3.10 Management Functions

The management and processing control of the large and complex student database was possible through the establishment and maintenance of the tracking file. Each record on this file contained the administration information for a single session, including absentee, excluded student, and assessed student counts. It also contained the processing history of that session's data, including the time, date, and number of booklets processed at the entry, verification, and resolution stages.

Using the Common Data Dictionary (CDD) product on the VAX, a domain was defined and stored for the tracking file, along with a corresponding record format, giving a label to each data field on the tracking record. Several procedures were developed using DATATRIEVE and stored in the CDD which accessed the tracking file and produced ad hoc processing status reports. Both procedures were provided to the operations coordinator for producing these reports.

The COUNTS procedure produced a summary of the various counter fields: number of students assessed, number of booklets entered, number of booklets verified, and number of booklets resolved. The ACTIVITY procedure produced a more detailed accounting of the counts at the session level, producing subtotals at the school and PSU level. The processing dates were also included to assist in the determination of any anomalies in the counts. DATATRIEVE was also used by the operations coordinator to make any corrections to the tracking file. A separate form containing all of the tracking record fields was developed and stored in the forms library. This form was linked to the file through the domain definition under CDD and processed via the DATATRIEVE modify command.

8.3.11 Data Spooling

At the completion of entry processing for the student database, the individual batch data and audit files were "spooled" into single, separate data files. In one step, this consolidation process accomplished three objectives: performing a final validation check on all data fields on all data records; preparing transfer of the database to the IBM mainframe; and facilitating the operation of quality control and descriptive analysis procedures.

The spooling program worked from the tracking file to ensure the processing of all batches. The resolution flag on each tracking record was checked to verify that resolution processing had been completed for that batch. Any unresolved batch was identified and noted by the program and processing continued with the next batch. The resolved batch data and audit files were opened for input processing. The program appended the session identification code to each input data record before writing it out to the spool data file. The audit records already contained session identification and were written to the spool audit file as is.

The spool data file is organized as an indexed file, using the session code and booklet serial number as the access key. The spool audit file is a sequential file.

An update program was made available to the operations coordinator for making corrections to the database after the fact of entry processing. This program operated inwardly and outwardly as the data entry program with the difference that the tracking file was not used and the data resided in one large file.

Chapter 8.4

EDITING DATA

Alfred M. Rogers

Educational Testing Service

The data editing process is divided into three separate steps: validation, identification, and correction. Validation ensures that each data value in the computer file is of the correct type, is within a range or set of ranges of values, and is consistent with other data values. All invalid data values are then identified and located in the raw data. The erroneous data are then either corrected or flagged as unresolvable in the computer file.

The errors uncovered by the editing process fall into two types: those made by the respondent (e.g., choosing two responses for a multiple choice exercise requiring only one response) and those made by data entry. The validation process reports both types of error with no knowledge of their source. The identification process determines the type of each error. The data entry errors are, for the most part, correctable; the correct value can be determined from an examination of the raw data. Errors made by the respondent, however, are difficult, if not impossible, to correct. If the intent of the respondent cannot be determined, the error must remain unresolved, but be flagged in some way to prevent incorrect interpretation in analysis and reporting procedures.

8.4.1 Student and Questionnaire Data

The data entry system served as the first line of defense against bad data. As described above, all data values were validated for type and range as they were entered from the data terminal keyboard. Special codes assigned for multiple and indeterminate responses were recorded and reported via the audit trail. The indeterminate values were later corrected under the resolution process.

At the completion of data entry processing, all of the batch student data files were "spooled" onto a single master file in preparation for transfer to the IBM mainframe. A second validation was performed during this spooling process to catch errors that had "slipped through" the entry system. An editing program was developed for applying corrections to this master file, using the same methodology as for the data entry program. This master file also served as the basis for preliminary descriptive data analyses and quality control checks.

Although the questionnaire files did not need to be spooled, they received the same secondary validation processing as the student data. Special attention was given to the "circle all that apply" items to ensure consistency in the coding of responses: if a respondent circled one or more of the alternatives, those would be coded "1" while the rest would be coded "0"; if no alternatives were marked, yet the respondent had the opportunity to reply, all fields would be coded "0"; if no alternatives were marked and the respondent had not reached the item or was instructed to skip it, all fields would be coded as "no response".

8.4.2 Professionally Scored Items

The professionally scored items went through a separate entry and editing process. The scoring of the items occurred after the booklets had been processed through the entry system. Since it was neither feasible nor economically prudent to send the booklets back through the entry system for just a few data values for each booklet, these items were processed by key entry systems.

The scores were entered by the raters into specially provided boxes on the back covers of the booklets. The boxes were arranged into rows for each of the items to be scored, with as many boxes in each row as there were scores permitted by the scoring guide for that item. Rather than devise a different format for each of the booklet types to be entered, a general-purpose format was implemented by allocating a maximum number of scores for each item and a maximum number of items per booklet. The scores were then keyed as a continuous string of data values into the separate item locations in each record. In addition to the scores, each record contained the student ID number for linkage with the master student file, and the rater ID codes from inside the front cover.

To ensure that the student ID codes were keyed accurately, the data file received from key entry was matched against the master file by the student ID, reporting any mis-matches from either file. The mis-matched ID codes were corrected on the input file and the matching program run again until there were no discrepancies.

The data files received from key entry were "loosely" formatted; the codes within the boxes were transcribed as a continuous string for as many rows as there were items in each booklet. Any processing scheme must use the booklet number within the student ID code to determine which items are in each booklet and the location and number of data fields to be processed for each item.

The validation program checked all the fields on each record for data type, range of values, and logical consistency with other fields. The rater ID fields were the first to be processed. The values of the ID codes were checked against a list of valid ID codes. The number of rater IDs was also noted for comparison with the number of scores per item; if, for any item, the number of scores disagreed with the number of raters, either a score value was missing or the rater ID code was not entered. For most of the booklets, only one rater performed the scoring. A 20 percent sample of booklets was selected throughout the scoring process and re-scored for reliability checking. These booklets would have two rater IDs. If for any item the first rater had disagreed with the second rater, the item was submitted to a scoring supervisor for resolution scoring. These booklets would have three rater IDs.

The program would then refer to a control table, indexed by booklet number, to determine the number and types of item score fields to process. The scored reading and writing items fell into five basic types: primary trait score only; primary and secondary trait scores; primary trait and holistic scores; primary, secondary and holistic scores; and one item, "The Door", which was subject to a mechanics scoring process. Twenty percent of the primary trait and holistic scores were subject to secondary and resolution scoring; the secondary trait items were scored only once. Processing the record continued on an item-by-item basis.

The primary trait scores, if applicable, were first checked for valid values according to the scoring guide, then counted for comparison with the number of rater IDs noted previously. If there were more than one score present, the values of the first and second scores were compared. If they disagreed, the program checked for the presence of both a third score and three rater IDs for the booklet. If they agreed, only two rater IDs were required.

The secondary trait scores, if applicable, were then validated according to the scoring guide. The program referred to the control table mentioned above for the number of secondary trait scores to be processed.

If the item was holistically scored, these scores were validated against the scoring guide, then counted for rater ID comparison. If more than one score was present, the values of the first and second scores were compared. If they agreed, only two scores and two rater IDs were required. If they disagreed by only one point, a third score was assigned by selecting the high or low value on an alternating basis throughout the execution of the program. If the scores disagreed by more than one point, the program checked for the presence of both a third score and third rater ID.

The validation program produced a printed list of all errors and inconsistencies found in the score file. The booklets identified in this list were collected and checked against the listing. In cases where a value had been mis-keyed, the correct value could be directly replaced in the data file. If, however, the error was on the booklet itself and accurately transcribed, the booklet was sent back to the scoring supervisor with an explanation of the error.

In either case, the data values on the file were corrected through the execution of an update program. This program used as input a "parameter card" file, each record of which indicated the identification code of the data file record to be altered, the field position within that record, and the value to be substituted. This approach not only guaranteed accurate and consistent correction of the data fields, but by its printed output provided a document of all changes made to the data file.

The corrected file was again processed by the validation program to ensure that all errors had been fixed and that no new problems were created by these corrections. If any more errors were found, the cycle of identifying the booklets, correcting the errors, and validating the corrected file was repeated until no more errors were found. At this point, the score file was ready for merging with the master student file.

8.4.3 Conclusion

Before the NAEP data entry methodology was developed, the editing process for any data file proceeded in the same manner as for the professionally scored items. The validation process was especially inefficient because it was performed after the fact of transcription and often by a second party who did not have immediate access to the raw data. Putting the validation mechanism at the point of entry removed most, if not all, of this inefficiency by informing the entry operator of a possible keying error while the raw data value was accessible. The interactive resolution process and audit trail mechanism also obviated the need to generate parameters for and run a generalized updating program as described above.

The editing process does not guarantee that all errors are removed from the data; only that the invalid, inconsistent, or otherwise unreasonable values have been at least identified, if not corrected. If a data value has been mis-keyed during the entry process and meets the validation criteria, this error could persist through the editing process to the analysis stage without detection. The verification process detects most of these errors by comparing independent entries of the same data and reporting discrepancies. The likelihood of an error surviving verification is thus very small, but still present. A quality control process must follow the entry and editing processes to ensure that the data values in a given record agree with the responses in the corresponding instrument.

Chapter 8.5

QUALITY CONTROL

John J. Ferris

Educational Testing Service

The purpose of quality control was to assess the accuracy of the data entry operation, or how closely the contents of the various instruments matched the resulting datasets. Even though the data were carefully keyed, verified, and edited, the question remains of how successfully this was done.

Whereas the editing operation assessed the data itself, the quality control operation assessed the process of entering the data. In editing, data records were selected (because inconsistencies were discovered by an editing program) and matched to the corresponding booklet; in the quality control work, the reverse operation was performed--booklets were selected and matched to the corresponding data record.

The examination of data records in the editing operation allows us to find some of the errors in all of the records; the detailed comparison between instrument and data record in quality control allows us to find all of the errors in some of the records. We cannot remove all errors from such a large and complex database as we have in NAEP. If an error has been made in key entry which appears sensible or reasonable in the data record, we cannot know it is an error unless that instrument happens to have been selected for quality control. That is why both editing and quality control are needed. Quality control allows us to discover potentially consistent problems in data entry which would never be discovered by an editing program. It also allows us to discover whether a database probably contains sufficiently valid data to support the analyses we wish to pursue.

Random booklets were selected and the actual instruments were compared keystroke for keystroke with the datasets created by the key entry system to discover the discrepancies and measure the quality of the data entry process. Overall, a very high quality was maintained throughout; the details are discussed below. The reader may wish to refer to data layouts or the instruments themselves in reviewing these details, especially when specific items are mentioned.

8.5.1 The Student Data

One of each booklet for each grade/age level was selected for analysis. Thus, a total of 67 booklets times three grade/age levels, or 201 booklets, were examined. A total of 111,421 keystrokes was involved in these 201 booklets; only 2 keystrokes were in error. This is an error rate of .000018.

However, since these results are affected by the chance selection of booklets, a further calculation was made. The probability of finding two errors in a sample of 111,421 keystrokes when the true error rate is, say, .0001 is

$$\binom{111421}{2} \times .0001^2 \times (1-.0001)^{111419} = .0009$$

The corresponding probability of finding one such error is .0002; the probability of finding zero such errors is .00001. These values must be added to the .0009 for the probability of finding two or fewer errors. In other words, we can be 99.89 percent sure that the true error rate for the student data booklet key entry operation was less than or equal to .0001.

8.5.2 The Excluded Student Questionnaire Data

Throughout the entire series of questionnaires in the NAEP database, a recurring problem was the treatment of multiple responses made to questions designed for a single response. An attempt was made at an early stage in the data entry to accommodate these multiple responses by extending the response code list to include codes for the multiple responses encountered. Inevitably, subsequent checking discovered the need for still more of these additional codes or an occasional misuse of a previously defined one. The Excluded Student Questionnaire was no exception in this regard. A list of these additional codes may be found in the codebooks accompanying the *NAEP 1983-84 Public-Use Data Tape Version 3.1 Users' Guide* (Barone, Norris, & Rogers).

Excluded Student Questionnaires were randomly sampled at the rate of 2.5 percent, or one booklet out of 40.

At Grade 4/Age 9,	58	booklets	checked	out	of	2354	
At Grade 8/Age 13,	56	"	"	"	"	2078	
At Grade 11/Age 17,	85	"	"	"	"	3485	
	199					7917	(2.514%)

The following discoveries and changes resulted from this process:

Grade 4/Age 9:

Multiple response resolutions were required for Question 15. Three new codes were added for this question, bringing the total number of codes to twelve, namely A-L. A remaining problem is that if the multiple response is B+C+E, it is recorded as F, the code for B+C. Other than this, 4 keystrokes were found to be in error.

Grade 8/Age 13:

Question 15 required similar attention at this grade/age. In addition, a number of questions with open-ended response alternatives were keyed without the corresponding response code because the respondent had neglected to circle it. The result could have been the loss of the write-in response if a database user were looking for the response code instead of the write-in response itself; accordingly, all of these response codes were added to the dataset. A total of 209 questions were affected. In addition, three keystrokes were found to be in error.

Grade 11/Age 17:

Other than re-coding of multiple responses as noted above, only one keystroke was found in error.

A total of 39,800 keystrokes was involved in the sample of Excluded Student Questionnaires examined. Disregarding the improvements in multiple response coding and the response codes added to 209 booklets at Grade 8/Age 13, there were actually only eight keystrokes in error. Applying the same analysis of this error rate as applied above, we can be 99.78 percent sure that the true error rate was less than or equal to .0005. Although this does not meet the standard set by the student data entry operation, it is also very reassuring.

8.5.3 The Teacher Questionnaire Data

As discussed above, this questionnaire also exhibited a shortage of special codes to reflect multiple responses which were far more common than had been anticipated. The lists of multiple response codes were expanded for a number of items and the additions were implemented in all booklets; these codes are defined in the codebooks accompanying the Public-Use Data Tapes Users' Guide.

Teacher Questionnaires were randomly sampled at the rate of 3 percent, or one booklet out of 33.

At Grade 4/Age 9,	26	booklets checked out of	1030	
At Grade 8/Age 13,	25	" " " "	791	
At Grade 11/Age 17	30	" " " "	914	
	81		2735	(2.962%)

The following discoveries and changes resulted from this process:

Grade 4/Age 9:

Zeros and dashes were found to be used in an inconsistent manner; although not an error as such, to avoid possible confusion it was decided to make all booklets conform to a consistent standard: when a respondent reached an item of the "circle-all-that-apply" type, a zero was used to mean an alternative did not apply; when such an item was not reached or should not have been answered, a dash was used to indicate missing data. One booklet had a write-in response which was re-interpreted. Other than these data adjustments, only three keystrokes were found to be in error.

Grade 8/Age 13:

The zero/dash confusion was found in some booklets and changed. As in the Excluded Student Questionnaire for Grade 8/Age 13, a number of questions with open-ended response alternatives were keyed without the corresponding response code because the respondent had neglected to circle it; these response codes were added. Three items, 21, 22, and 23, were lacking a response flag position in the booklet though one had been provided in the data layout; since the codes had therefore not been keyed, they were added by program. Seventeen erroneous keystrokes were found.

Grade 11/Age 17:

The zero/dash confusion was found in some booklets and changed. Four keystrokes were found to be in error.

A total of 41,398 keystrokes was involved in this sample of Teacher Questionnaires. Twenty-four of these keystrokes were in error. The application of the above-described error analysis allows us to say that we are 99.76 percent sure that the error rate is less than or equal to .0010. This rate is twice as high as that found for the Excluded Student Questionnaire and ten times as high as that found for the student data.

Although this error rate is perhaps not alarmingly high (it suggests 99.9 percent "pure" data), it does reflect a characteristic of the Teacher Questionnaires that was observed during quality control and editing operations: this instrument seemed to be unexpectedly difficult for teachers. Again and again strange answers, inconsistent answers, missing answers and mis-answered questions were found throughout the data. Perhaps this explains the relative difficulty of keying this data correctly.

8.5.4 The School Characteristics and Policy Questionnaire Data

This questionnaire suffered somewhat for its design and the quality of the responses. Two items, write-ins dealing with reading programs, could not be dealt with meaningfully. A number of questions asking for percents were answered unpredictably: N's may have been used instead of percents; percents or proportions may have been used instead of N's; percents were indicated but did not add up to 100 proportions were confused with percents. Some write-in responses were too long to be accommodated in the fields provided in the database; such responses can only serve a flagging purpose. Also, many of the same sorts of problems were encountered here as were encountered with the Teacher Questionnaires.

School Characteristics and Policy Questionnaires were randomly sampled at the rate of 5 percent, or one booklet out of 20.

At Grade 4/Age 9,	30	booklets	checked	out	of	623	
At Grade 8/Age 13,	27	"	"	"	"	459	
At Grade 11/Age 17,	15	"	"	"	"	301	
	72					1383	(5.206%)

The following discoveries and changes resulted from this process:

Grade 4/Age 9:

The zero/dash confusion described above was encountered with some frequency. Some additional multiple response codes were added. Fifteen keystrokes were judged to be in error.

Grade 8/Age 13:

The zero/dash confusion described above was encountered with some frequency. Some additional multiple response codes were added. Ten keystrokes were judged to be in error.

Grade 11/Age 17:

> The zero/dash confusion described above was encountered with some frequency. Some additional multiple response codes were added. Only one keystroke was in error.

A total of 31,536 keystrokes was involved in this sample of School Characteristics and Policy Questionnaires. With 26 keystroke errors, we can be 99.78 percent sure that the true error rate is less than or equal to .0014. Some of the factors contributing to this error rate have been noted above. Again, the complexity of the instrument and the occasionally careless manner in which some of the questions were answered certainly added to the difficulty of the keying operation. While this error rate does not meet the extremely high standard set by the data entry for the student data, it does indicate a level of excellence seldom encountered in a database of this size.

8.5.5 Summary of Error Analysis

The quality control of the NAEP data for Year 15 revealed very high standards of data entry for all instruments. In the interests of making the data easier to interpret and preserving more of the complexity of the data, some changes were made which were considered improvements rather than correction of errors. The errors that were discovered led to the following assessments of likely error rates.

	Error Rate	Confidence True Rate is < or =
Student Data	.0001	99.89%
Excluded Student Questionnaire	.0005	99.78
Teacher Questionnaire	.0010	99.76
School Characteristics Questionnaire	.0014	99.78

Chapter 8.6

DATABASE CREATION

Alfred M. Rogers

Educational Testing Service

The data transcription and editing procedures described in Chapter 8.1 resulted in the generation of disk and tape files containing various assessment information. Before any analysis could begin, these files had to be pulled together into a comprehensive, integrated database. Sampling weights were also required in order to make any valid statistical inferences about the population from which the assessment sample was drawn.

This chapter describes the processes of extraction of sample information for the derivation of sampling weights, and the merging, or bringing together, of the many transcription files into the NAEP database.

8.6.1 Extraction

For each grade/age cohort, four sets of weights were required to perform inferential analyses: school weights, excluded student weights, student weights, and teacher weights. Due to the method by which teachers were selected, sampling weights could not be assigned to teachers, but were instead assigned to students who were linked to participating teachers. (See Chapter 7 for more details.)

All of the sample information was extracted from the data files, edited, and transferred to tape files for shipment to Westat, where the weight computation was performed. The editing process included both the validation of the data values as well as frequency distribution analyses to be compared with tracking information from the data entry system.

The school sample information was available to Westat from the beginning of the assessment. They did not require any additional information from ETS to compute school sample weights.

The excluded student sample information was extracted from the Excluded Student Questionnaire data file. This information included: booklet serial number, PSU and school code, grade, sex, birth date, race/ethnicity, and a code indicating reason for exclusion. All data fields were taken from the front cover information of each booklet, except for the exclusion code, which was derived from the response to Item 3 of the questionnaire. A listing of the Excluded Student Questionnaires which had not been received at ETS was included with the file for each grade/age cohort.

The student sample information came from two sources: the student database and the absentee file from the administration schedules. The assessed student sample information included: booklet serial number, PSU and school code, grade, sex, birth date, race/ethnicity, and teacher code. Since the absent students were not observed and not assigned an assessment booklet, the booklet serial number, race/ethnicity, and teacher code were not available for the absentee data.

The absentee file had to be adjusted for makeup sessions. The field administration procedures required scheduling of makeup sessions if absentee rates exceeded certain limits. The students attending these makeup sessions were supposed to be originally sampled students who were absent for the regular sessions. Failure to remove the makeup students from the absentee file would have resulted in incorrect estimates of the number of students in those schools. This problem could have been particularly acute in the Grade 11/Age 17 sample where absentee rates were high and many schools required makeup sessions.

The first step in the removal process was to identify the students in the student file who attended makeup sessions in each school. Then, for each school and session type (spiral or tape), the sex, grade, and birth dates of the makeup students were matched with those of the absentee students in the same school and session type. The absentees identified by perfect matches were removed from the absentee file; the remaining unmatched makeup students, if any, were paired with randomly selected absentees who were then removed from the file. This latter procedure was necessary only for the Grade 11/Age 17 sample in only a few of the many schools which had makeup sessions.

The teacher sample information was extracted from the teacher questionnaire data file. It consisted of only the PSU, school, and teacher codes from the questionnaire booklet covers. Westat used this information in conjunction with the student sample information to produce a file of student-based teacher weights.

8.6.2 File Merging

The transcription process resulted in the generation of five data files for each grade/age cohort: one file for each of the three questionnaire instruments, the student response data file from the data entry system, and the student reading and writing scores from professional scoring and key entry. The sample weight derivation process produced an additional four files of sampling weights. To perform data analysis, these files had to be integrated into a coherent and comprehensive database.

This database would ultimately consist of four files per cohort: school, teacher, excluded student, and student files. The student file would contain all five student samples: the spiral and four tape samples.

The school file could be linked to the other three files through the PSU and school codes. The teacher file could be linked to the student spiral sample through the PSU, school and teacher codes.

The school file was created by merging the School Questionnaire file with the school weights file. The PSU and school code were used as the matching criterion. Each record of the resulting file was formed by concatenating the weight information with the response data. Since not all schools returned their questionnaires, some of the output records contained only weight information.

The teacher file was generated from the Teacher Questionnaire file. Since the teacher weights were derived at the student level, no information had to be added to the questionnaire data.

The excluded student file was the result of merging the Excluded Student Questionnaire file with the excluded student weights file. The booklet serial number was used as the matching criterion.

The creation of the student data file was a three-stage process, merging the professionally scored items, student weights, and teacher-based student weights with the student response data, in that order. In all three procedures, the booklet serial number was used as the matching criterion. The merging of the professionally scored item data was a more complex procedure than the others, because the set of scores for each item within a booklet were inserted into the response data fields in the order in which the items appeared in the booklet.

The database was then ready for analysis. As new data values and scores were derived, they were added to the relevant files using the same matching procedures as described above. The public-use data tapes files were ultimately generated from this database.

8.6.3 Master Catalog

A critical part of any database is the processing control and descriptive information. A central repository of this information may be accessed by all analysis and reporting programs to provide correct parameters for processing the data fields as well as consistent identification labeling of the analysis results. The master catalog file was designed and constructed to serve both of these purposes.

Each record of the master catalog contains the processing, labeling, classification, and location information for each data field in the database. The control parameters are used by the access routines in the analysis programs to define the manner in which the raw data values are to be transformed and processed.

All data fields have a 50-character label in the catalog describing the contents of the field and, where applicable, the source of the field. The data fields with discrete or categorical values have additional label fields in the catalog containing the permitted values and 8- and 20-character labels for those values.

The classification area of the catalog record contains distinct fields corresponding to pre-defined classification categories for the data fields. For a given classification field, a non-blank value indicates the code within that classification category for the data field. This permits the collection of identically classified items or data fields by performing a selection process on one or more classification fields in the catalog.

According to the NAEP design, it is possible for item data fields to occur in more than one age assessment and more than one block within each age. The location fields of the catalog record contain the age, block and, where applicable, the item sequence number within block of each occurrence of the data field throughout the Year 15 database.

The master catalog file was constructed in parallel with the collection and transcription of the assessment data to be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their descriptive and control information was entered into the catalog.

One of the most important uses of the master catalog was the control of the creation of the public-use data tapes files as well as the codebooks and file layouts. A synopsis of this process is presented in the next chapter.

Chapter 8.7

PUBLIC-USE DATA TAPE CONSTRUCTION

Alfred M. Rogers

Educational Testing Service

The public-use data tapes (PUDTs) are designed to permit any research individual or organization with an interest in the National Assessment to perform secondary analysis on the same data as that used at ETS. This section discusses some of the issues raised during the creation of the data, and summarizes the procedures followed in generating the data and related materials.

The three elements of the distribution package are the data tapes, printed documentation, and microfiche of the assessment instruments. Each grade/age cohort is represented on a separate tape, with each tape containing the data files; a set of SPSS-X control statement files for generating an SPSS-X system file for each data file; a set of SAS control statement files for generating a SAS system file for each data file; and a set of machine-readable catalog files containing control and descriptive information for each data file, for the non-SPSS-X and non-SAS user. The printed documentation consists of four volumes: a guide to the use of the data files, and a set of file layouts and codebooks for the data files within each of the three cohorts (see NAEP 1983-84 Public-Use Data Tapes Version 3.1 Users' Guide [Barone, Norris, & Rogers, 1986]).

8.7.1 File Definition

The organization and format of the data files to be produced was the first issue to be addressed. The ETS database consisted of four data files for each grade/age cohort, corresponding to the three questionnaire instruments and the student database, incorporating the spiral and all four tape samples. The logical relationship of the data files was a three-level hierarchy, with the five student and the excluded student samples at the bottom level; the teacher sample at the next level, with a linkage only to the spiral sample; and the school sample at the top, with direct linkages to all samples below it. A linkage may be viewed as a one-to-many mapping of the records within the two files linked. For example, one school record is linked to one or more records in the teacher file, and each of these teacher records are in turn linked to one or more records in the spiral student file.

One organization scheme has six files corresponding to the six samples at the bottom level, with the data from the higher order samples appended to and repeated across as many of the lower level records as required by the linkages. Using the previous example, each spiral sample record would be appended by its corresponding teacher record and school record. This approach places no demand on the user to define the linkages since each data record is complete, but it requires substantially more computer storage space due to the larger record size.

An alternative scheme would have these same six samples without the appended teacher and school data. The teacher and school samples would reside in their own files, with special data fields in all files to facilitate their linkage through program control. At the expense of a little more sophistication on the part of the user, this approach is more economical in computer resource utilization. This potential for savings on computer storage and processing costs was the overriding consideration in using this scheme.

8.7.2 Variable Definition

The selection and arrangement of variables, or data fields, in each file was the next order of business. The first step in the decision process was the generation of a file of variable descriptors for each data file to be created. Each of these LABELS files contained one record for each variable, each record containing the variable name, a short text description of the variable, and processing control information to be used by later steps in the PUDT process. This file could be edited for deletion of variables, modification of control parameters, or re-ordering of the variables within the file.

The first program in the processing stream, GENLYT, produced a printed layout for each file from the information in its corresponding LABELS file. These layouts were initially reviewed for the selection and ordering of the variables. The variables which were excluded from PUDT processing fell primarily into two categories: non-applicable and confidential.

The non-applicable variables were found mostly in the student database. Since the tape samples were combined with the spiral sample, many of the variables which applied to the spiral students did not apply to the tape students, and vice versa. For example, the teacher code and the student-based teacher weights were used for the analysis of spiral sample data, but were not in the design at all for the tape sample.

The confidential variables included any descriptor or code which could be used to identify individual states, schools, or students in the NAEP sample. The PSU, school, teacher, and student identification codes used internally by ETS and WESTAT were "scrambled" according to specific algorithms to obtain new codes for use in linking the files together.

Another confidentiality problem arose in the response data, where the students were asked to identify the state they had lived in four years ago.

A new variable was created using the response code and current state residency information from the PSU code to indicate if the student had lived in the same state, the same region, or a different region.

The ordering of the variables within the data files followed a general trend of decreasing likelihood of usage: identification information preceded weights, scores, and other derived variables, which were followed by the response data. The identification variables were generally those on the front covers of the instruments. The derived variables included the sampling weights, IRT scale values, and variables derived from the response data or other sources for reporting purposes. The response data variables were arranged according to their order in the instrument.

The spiral sample posed an additional problem because it entailed the expression of 63 different booklet formats into a single, fixed format. The solution lay in arranging the data "blocks" in order corresponding to their letter designations. The common background questionnaire preceded the first spiral block in the new record. Each data record from the input student base was reformatted according to its booklet number; the data for its constituent blocks were moved into their assigned locations in the output record. The remaining data block areas contained blank fields, indicating that the data was missing by design.

The spiral design also created a problem from the user's standpoint: how to determine, from a given booklet record, which data blocks were present and their relative order in the instrument. This problem was remedied by the creation of a set of control variables, one for each block, which indicated not only the presence or absence of the block but its order in the instrument. These control variables were included in the section of derived variables.

8.7.3 Data Definition

To enable the data files to be processed on any computer system using any procedural or programming language, it was desirable that the data be expressed in numeric format. With the exception of a handful of open-ended responses, this was possible, but not without the adoption of certain conventions for re-expressing the data values.

As mentioned in Chapter 8.3, Data Entry, the responses to all multiple-choice items were transcribed and stored in the database using the letter codes printed in the instruments. This scheme afforded the advantage of saving storage space for items with ten or more response options, but at the expense of translating these codes into their numeric equivalents for analysis purposes. The response data fields for most of these items would require a simple alphabetic-to-numeric conversion. However, the data fields items with ten or more response choices would require "expansion" before the conversion, since the numeric value would require two column positions. One of the processing control parameters on the LABELS file indicates whether or not the data field is to be expanded before conversion and output.

The ETS database contained special codes to indicate certain response conditions: no response, "I don't know" response, multiple response, and unresolvable response. The primary trait scores for the reading essay items included additional special codes for ratings of "illegible" and "off task" by the scorers. A final special code was assigned to the items which, due to printing error, did not appear in some of the booklets at all. These codes had to be re-expressed in numeric format.

A convention used by ECS in the creation of their Public-Use Data Tapes was adopted and enhanced in the designation of these codes. The "I don't know" response was always coded as 7. The "no response" code was 8. The multiple or otherwise indeterminate response received a code of 9. For the primary trait scores, an "illegible" score was coded as 5 and the "off task" score as 6. The very small number of "missing" responses were coded as blank fields, corresponding to a "missing by design" designation.

This coding scheme created conflicts for those items which had seven or more valid responses as well as the "I don't know" response, and the primary trait items with five or more scoring categories. These items also required expansion to accommodate the valid responses values. The special codes were "extended" to fill the output data field, e.g. the "I don't know" code was extended from 7 to 77, the "no response" code from 8 to 88, etc.

The numeric variables on the tape files fall into two categories: continuous and discrete. The continuous variables include the weights, IRT values, identification codes, and item responses where counts or percentages were requested. The discrete variables include those items for which each numeric value corresponds to a response category. This designation also includes those derived variables to which numeric classification categories have been assigned. The open-ended short response items were to be transferred with no conversion, and were classified as alpha-type variables.

8.7.4 Data File Layouts

The data file layouts, as mentioned above, were the first user product to be generated in the PUDT process. The generation program, GENLYT, used a LABELS file as input and produced a printable file. This LAYOUT file is little more than a formatted listing of the LABELS file.

Each line of the LAYOUT file contains the following information for a single data field: sequence number, field name, output column position, field width, number of decimal places, data type, value range, key or correct response value, and a short description of the field. The sequence number of each field is implied from its order on the LABELS file. The field name is an 8-character label for the field which is to be used consistently by all PUDT materials to refer to that field on that file. The output column position is the relative location of the beginning of that field on each record for that file, using bytes or characters as the unit of measure. The field width indicates the number of columns used in representing the data values for a field. If the field contains continuous numeric data, the number of decimal places value indicates how many places to shift the decimal point before processing data values.

The data type category uses three codes to designate the nature of the data in the field: alpha-numeric data are coded "A"; continuous numeric data are coded "C"; discrete numeric data are coded "D". Additionally, the discrete numeric fields which include "I don't know" response codes are coded "DI". If the field type is discrete numeric, the value range is listed as the minimum and maximum permitted values separated by a hyphen to indicate range. If the field is a scorable item response, the correct response value, or key, is printed. A range of correct responses was indicated for those professionally scored items which received cut-point scoring for IRT scaling. Finally, each variable was further identified by a 50-character descriptor.

8.7.5 Data File Catalogs

The LABELS file contains sufficient descriptive information for generating a brief layout of the data file. However, to generate a complete codebook document, substantially more information about the data is required. This function is filled, in part, by the CATALOG file.

The CATALOG file is created by the CATGEN program from the LABELS file and the Year 15 master catalog file. Each record on the LABELS file generates a CATALOG record by first retrieving the master catalog record corresponding to the field name. The master catalog record contains usage, classification, and response code information. This record is prefixed by the positional information from the LABELS file: field sequence number, output column position, and field width.

The response code information, also referred to as "foils", consists of the possible data values for the discrete numeric fields, and a 20-character description of each one. The CATGEN program uses additional control information from the LABELS file to determine if extra foils should be generated and saved with each CATALOG record. The first control parameter or "flag" indicates a primary trait score field, for which the "illegible" and "off task" codes and foils should be generated. The second flag controls generation of the "I don't know" foil. The third flag regulates "no response" foil generation, and the fourth flag denotes the possibility of multiple or out-of-range responses for that field and sets up an appropriate foil. All of these control parameters, including the expansion flag, may be altered in the LABELS file by use of a text editor to suit the data behavior for any given field.

The LABELS file supplies control information for many of the subsequent PUDT processing steps. The CATALOG file provides the detail information for those same steps and for others as well.

8.7.6 Codebooks

The data file codebook is designed as a printed document containing complete descriptive information for each data field. Most of this information derives from the CATALOG file; the remaining data came from two other files: the COUNTS file and the IRT parameters file.

Each data field receives at least one line of descriptive information in the codebook. If the data type is either alphabetic or continuous numeric, no more detail is given. If the variable is discrete numeric, the codebook lists the foil codes, foil labels, and frequencies of each value in the data file. Additionally, if the field represents an item used in IRT scaling, the codebook lists the parameters used by the scaling program.

The frequency counts are not available on the catalog file, but must be generated from the data itself. The GENFREQ program created the COUNTS file using the field name to locate the variable in the database, and the foil values to validate the range of data values for each field. This program also serves as a check on the completeness of the foils in the CATALOG file, as it flags any data values not represented by a foil value and label.

The IRT parameter file is linked to the CATALOG file through the field name. Printing of the IRT parameters is governed by a control flag in the classification section of the CATALOG record.

The LAYOUT and CODEBOOK files are written by their respective generation programs to print-image disk data files. Draft copies are printed and distributed for review before the production copy is generated. The production copy is printed on an IBM 3800 printing sub-system using laser-imaging technology. The printing is performed at 15 characters per horizontal inch (pitch) and 8 lines per vertical inch. This accommodates printing of 115 characters per line and 80 lines per page on standard 8-1/2" x 11" paper.

8.7.7 SAS and SPSS-X Control Files

The SAS and SPSS-X control statement files are provided to the user as a means for converting the raw data files directly into a system file for subsequent analyses under either package. The files are very similar in their content and structure, although actual implementation of their features differ slightly. Two separate programs, GENSAS and GENSPX generate the control files using the CATALOG file as input.

Each of the control files contain separate sections for variable definition, variable labeling, missing value declaration, value labeling, and creation of scored variables from the cognitive items. The variable definition section describes the locations of the fields, by name, in the file, and, if applicable, the number of decimal places or type of data. The variable label identifies each field with a 50-character description. The missing value section declares which values of which variables are to be treated as missing and excluded from analyses. The value labels correspond to the foils in the CATALOG file. The code values and their descriptors are listed for each discrete numeric variable. The scoring section is provided to permit the user to generate item score variables in addition to the item response variables.

Each of the code generation programs combine three steps into one complex procedure. As each CATALOG file record is read, it is broken into several component records according to the information to be used in each of the resultant sections. These record fragments are tagged with the field sequence number and a section sequence code. They are then sorted by section code and sequence number. Finally, the reorganized information is output in a structured format dictated by the syntax of the processing language.

The generation of the system files accomplishes the testing of these control statement files. These files are saved for use by internal ETS users of the NAEP data.

8.7.8 Machine-Readable Catalog Files

For those NAEP data users who have neither SAS nor SPSS-X, yet require processing control information in a computer-readable format, the distribution tape also contains machine-readable catalog (CAT) files. In addition to processing control information, each CAT record contains the IRT parameters and the foil codes and labels.

IMPLEMENTING THE NEW DESIGN: THE NAEP 1983-84 TECHNICAL REPORT

PART II

Chapter 9

OVERVIEW OF PART II:
THE NAEP 1983-84 DATA ANALYSIS

Albert E. Beaton

Educational Testing Service

The purpose of this chapter is to present an overview of the procedures used in the analysis of the NAEP Year 15 (1983-84) data. These procedures were used for the parameter estimates which were reported in The Reading Report Card: Progress Toward Excellence in Our Schools (1985), Writing: Trends Across the Decade, 1974-84 (Applebee, Langer, & Mullis, 1986), and other reports which have been prepared or are in preparation. The details of the analytic procedures are described in detail in the rest of Part II of this technical report.

This second part of the technical report assumes the existence of a carefully edited database, thus does not cover the operations involved in constructing the database, which are are discussed in Part I. The reader should consult Part I of this report for information about the NAEP data, including:

- an overview of the operations involved in collecting and editing the data (Chapter 2);
- the development of the reading and writing exercises (Chapter 3);
- the stratified random sampling plan (Chapter 4);
- the assignment of exercises to students (Chapter 5);
- instrument and item information (Chapter 6);
- the field administration procedures (Chapter 7); and
- the flow of data from the field to an edited database and public-use data tapes (Chapter 8).

Sections 9.1 through 9.6 below follow the sequence of the remaining chapters in Part II of the technical report:

- the reading data analysis, including the study of dimensionality, differences between printed and tape recorded administration, maximum likelihood estimation, marginal estimation, conditioning, plausible values, trend analysis, and behavioral anchoring (Chapter 10);

- the writing data analysis, including reliability, differences between printed and tape recorded administration, trend analysis, ARM scaling, conditioning, and plausible values (Chapter 11);
- the background and attitude data analysis, including the definition of reporting variables and WARM scaling procedure (Chapter 12);
- the estimation of population parameters, including sampling weights, estimation of sampling error, estimating measurement error, and the NAEP tabular results (Chapter 13); and
- some supplementary studies, including the validity of the NAEP data and the design effects for Year 15 NAEP (Chapter 14).

* * *

Before discussing the data analytic procedures, it may be useful to make some general comments about the NAEP data analysis. The NAEP data analyses were guided by a number of priorities: accuracy, interpretability, public-usefulness of the data, and timeliness of reporting. There were also a number of constraints such as keeping a student's burden under an hour, maintaining trends, collecting data in the schools within a few months of receiving the grant, and, of course, keeping within a very tight budget. These priorities and constraints often conflicted and required improvisation.

An example of conflict occurred during the assembly of test booklets. The ETS design called for random subsamples of students to be administered booklets of about 25 exercises each, a number sufficient to permit a reasonably precise estimate of reading proficiency from each sampled student. Within the tight transition and assessment timelines, however, the target of 25 exercises per pupil could not be met, and it was not possible to obtain precise estimates of proficiency for all students. Because it is population-level characteristics rather than individual student characteristics that are of interest in NAEP, the possibility of fulfilling NAEP's function remained open--but only if new techniques could be developed to produce estimates of population attributes directly, without the (now impossible) intermediate step of computing scores for everyone in the sample.

The major tool used in computing consistent estimates of group performance was plausible values, an adaptation by Mislevy (1985a) of a method of handling missing data originally proposed by Rubin (1977, 1978). The idea is that although we do not know an individual's proficiency, we can estimate a distribution of plausible scores for each individual, given the available data, and that this distribution represents both what we do know and what we do not know about the individual's proficiency. Using a random selection from the distribution of plausible values of each individual, it is possible to make consistent estimates of selected population parameters. The variation of results from one set of random selections to another is an estimator of the error due to imprecise measurement. In practice, we generate five plausible reading and writing values for each individual who was administered exercises in the area and then, to estimate measurement error variance, compute each parameter estimate five times.

This method of estimation does not in general give consistent results for all possible subpopulations, and will not unless the classification variables corresponding to the subpopulations are explicitly conditioned on in the process of creating the plausible values. We therefore conditioned on as many variables as our technology would allow, which were the major NAEP reporting variables (e.g., sex, race/ethnicity). Beaton, Mislevy, Kaplan and Sheehan (1986) demonstrated the process using data available on the SAT Public-Use Tape and showed the importance of conditioning on subpopulation membership. Since then, the possible biases involved in using unconditioned variables have been studied extensively. The results are reported in the next two chapters.

The development of the concept of plausible values for large scale surveys has had several side benefits. Already, we have been able to place the reading data from past assessments onto the Year 15 scale for trend analysis, where the data might otherwise have been too sparse for scaling. Since fewer exercises are required for group estimates, limited assessment time can be used to estimate several subscales in a learning area, thereby reducing dependence on the assumption of unidimensionality (see Zwick, Chapter 10.1). Perhaps most important is that the plausible values force an analyst to consider an important problem which is hidden in much survey research: the problem of fallible measurement.

All educational measurement, indeed all measurement, is subject to error, and this error affects the assessment of relationships which are made from the data. This phenomenon affects all analyses of educational survey data. We have not studied whether or not other national surveys have more or less measurement error than NAEP, but some such error surely exists. The method of plausible values is an attempt to improve estimation, given the fallibility of the data.

Let us consider an assessment design which might have been used instead of the present NAEP design. We might have administered a short test of reading and writing to all students, at least at a particular grade/age combination. The most obvious losses would be the broad coverage of the subject areas, the links to past assessments, and, unless the tests contained substantial overlap at different grades, the linking across grade/ages. The measurement of an individual's proficiency would still be imprecise; the amount of error would depend, among other things, on how many items were offered, how many items a student attempted, and the selection of the items themselves, especially if a serious ceiling and floor effect were present. Such a survey would offer each student the same number of items in a subject area, whereas the NAEP design offers some students many items and other students only a few. Neither the hypothetical survey design nor the NAEP design is assured of an adequate range of items nor can either assure that the students will respond to all of the items offered. We would expect that the measurement error from subject to subject would vary less in the simpler design than in the NAEP design, which has both quite precise and imprecise subject measurement. In both cases, ignoring the measurement error is done at the analyst's peril, since the error will result in biased results.

The method of plausible values can be used in either case to reduce the bias due to error of measurement. If the assumptions of the plausible value models are met, the bias in parameter estimates approaches zero as the sample size approaches infinity.

The measurement error problem, and several other data analytic problems, have been known for years and affect analyses of any educational data, including the analyses that we have done of NAEP data and the analyses that others may do using these data in the future. The question might arise as to whether these data--indeed any survey data--should ever be used at all. The value decisions involved in using imprecise data were well described twenty years ago in Equality of Educational Opportunity (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966), which is better known as the Coleman Report. The Technical Appendix[1] to Section 3.2 states on page 326:

> There are three central facts to be remembered throughout any analysis of the sort here conducted:
>
> 1. The measurement of either any single variable or any class of variables is at best partial and incomplete.
>
> 2. When two variables (or two sets of variables) are statistically associated, for reasons that may be either irrelevant or closely related to the study, an apparent relationship of another variable to one of them may result from an actual relationship of that variable to the other. (If this occurs, we are likely to speak of the first as a surrogate for the second, and to try to uncover the effect by studying the joint relationship of our response with both variables or sets of variables.)
>
> 3. Even if association of the variables we are studying with some "explanatory" variable is firmly established, this establishment cannot of itself

---

[1] The author remembers that Professor John Tukey wrote this section.

> > settle the question of causation (though strong evidence would be provided if the time order were known); either variable may "cause" the other, or both may share a common cause. In many cases, continuing studies of the development of these variables over time can untangle such a question of "What causes what?" In the present case, studies of change in achievement level could give more direct evidence than the present cross-sectional survey.
>
> *To neglect any of these central difficulties is to lay oneself open to very serious risk of error.* *Yet, to fail to use such evidence in making judgments and taking action is to lay oneself open to the often more serious dangers of unwarranted inaction, or of action based merely upon rumor and ill-founded opinion.* We must recognize and deal with the three difficulties, using care in interpretation. (emphasis added)

These comments are as relevant to the NAEP data analysis as to analyses performed twenty years ago. We have tried to ease the measurement error problem. The other problems, surrogate variables and causation, are still a matter of concern and are left to the judgment of the secondary analyst and reviewer.

Computing the best parameter estimates that our technology allows is costly in both computer expenses and conceptual complexity. In the next two chapters, the technology is made available to those who choose to do so, as we did. Ultimately, however, it is up to the secondary user to decide the level of accuracy he or she needs and can afford. To minimize cost, a secondary user might use one plausible value as if it were a test score and proceed to do standard analyses using available statistical systems. We do not recommend this approach because it will often lead to bias and to underestimates of sampling error, although these problems may be no worse than those incurred using the data from other complex educational surveys. Simple analyses may be sufficient for exploratory data analysis, but more complex methods may be necessary for better results, especially when high order interactions or partial regression coefficients are involved. The next several chapters contain a number of suggestions for improving the accuracy of results at modest increases in cost and complexity, and we will continue to develop methods of handling such data in the future.

The NAEP database contains a very large amount of information about a carefully selected national sample. Part I of this technical report reveals the care that has gone into making the database as clean as possible. We wish to encourage secondary users to make full use of this important informational resource.

9.1 The Reading Data Analysis

A major strategy in our approach to the reading data analysis was the introduction of scaling to reduce the available information into a manageable and interpretable form. The reading data of the Year 15 assessment are quite voluminous, with 340 exercises administered over three grade/age levels. Reporting the percentage passing each exercise for each age, grade, gender, race, ethnicity, region, etc., seemed to us too burdensome for the potential audiences of NAEP's reports, although such information is available for those who are interested. It also seemed to us that reporting the average of the percentages passing, even when reported for selected subsets of exercises, did not take full advantage of the information available in the data. We intended to investigate whether the data fell approximately on a single dimension, and, if so, to use item response theory (IRT) methods to form a scale. The scale would then be interpreted, using behavioral anchoring, to make fairly general statements about what students could and could not do.

The reading data analysis also had to explore the effect of changing the method of exercise administration from tape recorder to print. To examine the change, we conducted parallel assessments: one by tape recorder, as in the past, and one using a printed presentation. If comparison of the resultant data showed that effect of the change followed a regular pattern, we intended to equate the new and old methods of administration and develop single trend lines for all data back to the first reading assessment in 1970-71. If the data showed substantial irregularity between the two data sets, we would present the results separately, with one curve showing trend from 1970-71 to the 1983-84 tape administration and a distinct point representing the 1983-84 print administration, the beginning of the future trend line.

The development of the reading scale required some improvisation. The ETS proposal assumed that a block of reading items would include about twelve scalable reading exercises which would span a wide range of difficulty so that few students would be able to answer either none or all of the exercises correctly. Because many students would be given two or three reading blocks, there would be a large, random, subsample of students who responded to around 24 items. Twenty-four items is approximately the number of exercises usually suggested for estimating individual performance using the maximum likelihood method. However, we did not know all of the properties of the reading exercises at the time of proposal writing and, within the transition time constraint, we were not able to form blocks of exercises that met the "twelve exercises to a block with varying difficulty" criterion. Some of the blocks had fewer exercises, some students did not respond to all the exercises offered, many students were able to answer all exercises correctly, and many others scored less well than would be expected by chance. The total effect of these factors was that maximum likelihood estimates of reading proficiency were attainable for only a non-random subsample of NAEP students. To rectify this situation, marginal estimation procedures were used to estimate a distribution of plausible reading proficiency values for an individual.

This procedure resulted in much more complicated methods for estimating national parameters.

The activities in the reading data analysis are described below. It should be noted that these steps were not always performed serially but, whenever possible, in parallel. We moved in parallel to improve the timeliness of results, but usually at additional cost. For example, the study of dimensionality was in progress at the same time as the scaling. As we learned more about our data and methodology, we re-ran previous steps. In logical order, the major steps were as follows:

(1) The dimensionality of the spiral reading sample was explored. Unidimensionality is an important assumption underlying the scaling methods that were used. Several different methods of assessing the dimensionality were employed. The analysis showed a general consistency of the data with the concept of unidimensionality and no compelling reason to use more than one dimension. The results of the dimensionality study are presented by Zwick in Chapter 10.1.

(2) The spiral sample, which used printed instructions, was compared to the tape sample, which used a tape recorder for administration. For the tape sample, the instructions were read aloud on a tape recorder, but the actual reading exercises themselves, obviously, could not be read aloud, although the student was paced through the exercises, that is, told when to move on the the next exercise. For this reason, the tape sample is sometimes referred to as the paced sample. We expected the effect of using a tape recorder to be regular and small, and thus equatable, although not ignorable, and it was. The study of the differences between printed and tape-recorded administration is discussed by Mislevy in Chapter 10.3.

(3) The maximum likelihood estimates of the parameters for selected items from the spiral data were estimated using the LOGIST program (Wingersky, Barton, & Lord, 1982). First, the data were fitted for each grade/age sample separately, and then for all ages and grades combined. Investigation showed that the differences between item parameters computed over all grade/age samples and those computed separately for each were negligible. Individual reading scores were calculated for all students who were presented at least seventeen items. This estimation procedure is described in detail by Wingersky in Chapter 10.2.

(4) The properties of these estimates were explored and it was found that finite estimates of reading ability were not available for about 15 percent of the sample. Furthermore, the inability to compute a reading score was associated with various background and attitude questions; for example, different ethnic groups had different rates of inestimable

scores. To respond to this problem and to make reasonable population estimates from these reading scores, Winsorization was tried, but did not seem to produce satisfactory results. Winsorization is discussed in Chapter 10.2.

(5) Because of the problems with the maximum likelihood estimates, the parameters were re-calibrated by Bayesian procedures using the BILOG program (Mislevy & Bock, 1982). In this re-calibration, only subjects who were administered at least two blocks were used. The item parameters generated by BILOG were similar to the LOGIST parameters.

Using the appropriate item parameters and the students' responses to the reading exercises and several background questions, a distribution of plausible values was computed for each student. This distribution represents the uncertainty involved in estimating an individual's reading proficiency: if a student responded to many exercises, the variance of this distribution would be small; if the students responded to only a few exercises, the variance would be large. For each student who was presented at least one reading exercise, five plausible values were randomly selected from his or her distribution. The BILOG scaling is discussed by Mislevy and Sheehan in Chapter 10.3.

Since the item parameters were essentially the same for all three grade/age combinations, the single reading scale, spanning all ages and grades, was used.

(6) A separate item calibration was performed for the Year 15 exercises administered using a tape recorder with data from past assessments, which used the same administrative procedures. These samples were available only for age, not grade, samples. The data for each age were calibrated separately.

The estimated item parameters from the tape-administered sample were compared with those computed from the spiral data which were administered by print. The parameter estimates were reasonably similar and so the estimates from the tape administered sample were equated to those of the print sample by means of randomly equivalent spiral- and tape-administered samples of each age population. This item calibration is reported in detail by Mislevy and Sheehan in Chapter 10.4.

(7) Some reading data from the past assessments in 1970-71, 1974-75, and 1979-80, as well as the data from the tape administered sessions of the 1983-84 assessment, were calibrated together using the BILOG program. The public-use data tapes from past assessments were the source of the student responses to exercises. Since not all past data could be linked to the present assessment, only exercises that were

also used in 1983-84, and other exercises in the same packages, were used. A separate calibration was done for each grade/age. The details of this analysis are reported in Chapter 10.4.

Since the reading scale is not appropriate for examining the sub-areas (literal comprehension, inferential comprehension, and study skills) that past reading reports have carried, the trends were continued by using the same reporting method used in the past. To maintain comparability with the past, only the tape administered exercises and only age, not grade, eligible students were used from the 1983-84 data. Only items that had been used in several assessments were used. The details are reported in The Reading Report Card (1985).

(8) The plausible values were then transformed to the NAEP reading scale. The NAEP reading scale takes the form of an estimated true score on a hypothetical test with known properties. This hypothetical test has 500 open-ended items, and all item responses are assumed to follow the logistic model. The items all have equal discriminating power and their difficulties are equally spaced across, and somewhat beyond, the range of student performance. Scores on this test can range between 0 and 500.

Several points on the reading proficiency scale were anchored. The purpose of anchoring the scale was to enhance interpretability by reporting what the vast majority of students at one level could do that most of the students at lower levels could not. Several scale points were chosen for anchoring: 150, 200, 250, 300, and 350. Reading exercises were chosen that discriminated between these reading score levels; the rule for exercise selection was that at least 80 percent of the students at one level could answer the exercise correctly whereas less than 50 percent at the next lower level could. The selected exercises were referred to a committee of reading specialists for interpretation. The result was verbal descriptions of the levels of reading performance. The scale transformation and behavioral anchoring is described by Beaton in Chapter 10.5.

(9) The performance levels of students at ages 9, 13, and 17 were then computed. Of particular importance was the percent of students who could read at or above the anchor levels. These percents, and all other reported statistics, were computed using the plausible values for reading. Standard errors were computed using the jackknife method. The methodology for estimating standard errors is discussed in Chapter 13.2.

9.2 The Writing Data Analysis

The writing data analysis involved (1) an initial investigation as to whether or not the writing data could support scaling, (2) the actual scaling, and (3) analysis of the scaled scores. The ETS proposal for the NAEP did not propose to spiral the writing data with the reading data nor to scale writing, and we did not intend to. However, the advantages of a summary scale that simplified age-to-age comparisons and facilitated secondary analyses led us to investigate the possibility of developing a writing scale, and we did. The result was a simple, useful scaling procedure and plausible values for a writing scale.

Scaling writing was quite different from scaling reading. First, the writing exercises were all professionally scored and assigned a rating between zero and four, whereas the reading data was in the form of, or could be converted to, right/wrong responses. Second, there were only 22 writing exercises, as compared to 340 reading exercises (of which 228 were used in the reading scale), and only ten exercises were spiralled so that their inter-correlations could be estimated. Finally, most students responded to only one or two exercises, and no one was administered more than four exercises. Thus, because of the non-binary nature of the exercises and the limitation on the amount of individual information, the well-developed item response theory that was applicable for reading proved inappropriate for writing.

The writing scale that we used is an extension of an idea presented by Goldstein and James (1983). The object of the assessment was to provide an estimate of how the population of students would have done, on the average, if all students were administered all of the ten essays that were spiralled. To reach this goal, we used the information available from a student's actual responses to estimate how he or she would have done if administered all essays. The estimates for individuals involved some uncertainty which was incorporated into the plausible values for writing and thus, ultimately, into the estimates of standard errors.

The steps in the writing analysis were performed in parallel wherever possible, as in the reading analysis. The major steps were as follows:

- Examining the rater reliability and computing basic statistics of the writing data. A random sample of 20 percent of the papers were scored twice by independent graders. The scorer reliability was computed separately for various essays and grade/age combinations. The average reliability was about .90. The means, standard deviations, and inter-correlations among the spiralled essays and selected background and attitude questions were computed for each grade. The results are discussed by Beaton in Chapter 11.1.

- Comparing the spiral and tape results. In writing, unlike reading, the actual question, as well as the assessment instructions, could be read to the students in the tape sample. Substantial differences in the distributions of item responses were found between those students who were administered the exercises using a tape recorder and those who were required to read the question. The details are presented by Johnson in Chapter 11.2.

- Analyzing the trend data. Since we did not feel that we could equate the spiral and tape results, we used only the tape results in analyzing trends. There were only a few essay exercises that were used in the past and offered in the tape sessions, and these were analyzed individually to produce the trend report. The details are presented by Johnson in Chapter 11.3.

- Scaling the writing data. Some of the essay exercises were administered at several age and grade levels, and the same scoring protocols were used, regardless of the ages or grades. The inter-correlations at the three grades were compared and found to be not significantly different. The three inter-correlation matrices were then pooled to make one correlation matrix.

  Using this correlation matrix and the responses of an individual student, an estimate of that student's average performance on all ten essays, and its standard error, was made. Assuming a normal distribution of error, five random values were selected from this distribution of plausible scores for that student.

  The writing scale can be labeled in the same way as the essays that it contains. The descriptions of levels of proficiency were the same for all essays; there were five ordered categories: a zero was no response, one was unsatisfactory, two was minimal, three was satisfactory, and four was elaborated. The common labeling for exercise responses automatically gave us a labels for scale points, but we found this implicit anchoring to be unhelpful, since the scale scores had a substantially smaller variance than the individual essays; thus, no students in the sample had scale scores of four. The details are presented by Beaton and Johnson in Chapter 11.4.

- For the cross-sectional report The Writing Report Card (Applebee, Langer, & Mullis, 19[illegible]6b), the average values for the different grade groupings and for demographic groupings were computed. Results were computed for each plausible value and the average result was used for reporting. All results were reported with their standard errors, which were computed using the jackknife method. The methods are discussed in Chapter 13.2.

9.3 Background and Attitude Analyses

Analysis of the background and attitude data has been largely restricted to the basic variables used in the report. Since trend analyses are restricted to variables used over time, these reporting variables are those used by ECS in past assessments. The detailed definitions are described in Chapter 12.

The racial/ethnic categorization has resulted in some detailed study which has been reported by Rivera and Pennock-Roman (1985). Since the first assessment in 1969, NAEP has asked its administrators to note the races of the students on the student listing form, hence the variable called "observed ethnicity." At first, students were classified only as black or white; in Year 3 (1971-72), the Hispanic classification was also observed. However, the small sample size for Hispanics precluded the creation of a separate reporting category until Year 11 (1979-80). In Year 7 (1975-76), NAEP began to ask each Age 17 student to report his or her own race or ethnicity, hence we also have "self-reported ethnicity." Self-reporting of race/ethnicity was added for Age 13 in Year 11 (1979-80) and for Grade 4/Age 9 in Year 15 (1983-84). After extensive study of the differences between the two definitions of race/ethnicity, the observed and self-reported race/ethnicities were combined into "imputed race/ethnicity." For trend reports, observed ethnicity was used, as in the past.

For The Writing Report Card, the background and attitude questions pertaining to writing were scaled using a variation of the ARM method, which was used for the writing exercise data.

The definitions of the background and attitude variables are discussed in Chapter 12.

9.4 Parameter Estimation

After the reading, writing, and background and attitude data were readied, the estimation of the competencies of students in American schools began.

The programs for parameter estimation, as well as many of the programs for data base creation and the analysis of reading and writing data, were written in F4STAT, the ETS proprietary statistical system (see Beaton, 1964; Beaton, 1973; and Educational Testing Service, 1984). F4STAT is a system of procedures for use with FORTRAN programs; the procedures include subroutines for data input and handling, matrix manipulation, statistical estimation, probability calculations, and output formatting as well as many other general purpose service routines. The procedures are assembled in an efficient manner for specific data analytic purposes. Although most, if not all, of the calculations done here could be computed using publicly available software, their costs and demands for machine resources might make these calculations prohibitively expensive.

First, the sampling weights were computed by Westat. In Chapter 13.1, Johnson, Hansen, Tepping, Lago, and Burke describe in detail how the weights were computed. Sampling weights were initially derived from the sampling design, then adjusted to account for nonresponse and trimmed to reduce sampling variance. Then, NAEP, Current Population Survey (CPS), and Census estimates of population sizes were combined into an optimum estimate of size for a number of subpopulations. Weights were computed for students, the teachers of these students, and schools.

Next, parameters for the nation as a whole and for specified subpopulations were estimated. The jackknife method was used to estimate the sampling error of the parameter estimates. Thirty-two synthetic samples were formed from the 64 PSUs in the sampling design, and separate student weights were computed for each of those synthetic samples. The original sample weight was used for parameter estimation and the synthetic samples were used for estimating sampling error. The details are covered in Chapter 13.2.

Another form of variability in parameter estimation comes from the uncertainty involved in the imputation of plausible values. Mislevy discusses this uncertainty, its estimation, and the use of plausible values in Chapter 13.3.

Finally, although many different statistics have been computed for various reports, certain simple, basic statistics are of such general value that we have computed them routinely and made them available to the NAEP staff for exploration, interpretation, and reference. Tabulating these simple statistics has resulted in many volumes of statistical tables which are sometimes referred to as almanacs. Tables cover both Year 15 and trend data. In Chapter 13.4, Zwick describes the basic tables, their contents, and their use.

9.5 Supplementary Studies

The Year 15 data analysis has entailed several supplementary studies which are reported here.

The ETS Standards for Quality and Fairness requires, among other things, the study of the validity of reported results. Applying the usual methods used for individual testing when the results are used only for groups of persons is inappropriate. Zwick describes how the content and construct validity of the reading and writing data were evaluated in Chapter 14.1. The study showed that, in general, the content- and construct-related evidence were supportive of the validity of the Year 15 NAEP reading and writing assessments.

Because using the jackknife is somewhat cumbersome for secondary analysts, we computed design effects for a number of parameter estimates. The design effect is a measure of the difference in efficiency in parameter estimation between a complex sampling design and a simple random sample, and can be used to simplify analysis procedures by achieving approximate results. Note that we have used the jackknife, not design effects, in all NAEP analyses; the design effects are for secondary analysis. In our opinion, the NAEP design effects were found to be reasonably small for this type of survey. Johnson provides the details in Chapter 14.2.

Chapter 10

THE READING DATA ANALYSIS: INTRODUCTION

Robert J. Mislevy

Educational Testing Service

This chapter describes the analyses carried out on responses to cognitive exercises in the Year 15 NAEP reading assessment leading to the results that appear in The Reading Report Card: Progress Toward Excellence in Our Schools (1985). The emphasis is on item response theoretic (IRT) scaling procedures, an innovation to NAEP beginning with the learning area of reading in the Year 15 assessment. This introductory chapter outlines general arguments for scaling, and discusses the special challenges that arise in the NAEP setting. Subsequent chapters detail the methods and results of specific procedures. Brief summaries of these chapters follow.

Chapter 10.1 - Dimensionality of cognitive reading exercises

It is a strong assumption to posit an IRT model in which a single examinee characteristic explains for responses to all items. This chapter describes analyses performed on the Year 15 BIB spiral data that explored the extent to which this assumption of unidimensionality is satisfied for the items included in the reading scale.

Chapter 10.2 - Joint estimation procedures

The ETS proposal for the analysis of the Year 15 reading data specified procedures based on the estimation of proficiency variables for each respondent. This chapter describes the analyses performed to this end, and documents the evidence for concluding the results were unsatisfactory.

Chapter 10.3 - Marginal estimation procedures

Alternative IRT procedures that do not require precise point estimates of individual examinee parameters are described in this section. These procedures include marginal estimation of item parameters and population characteristics, and "plausible values" associated with, but not estimates of, individual examinees' proficiencies. Also described here is the equating of the responses gathered under BIB spiral conditions in Year 15 to those gathered under paced conditions in previous NAEP assessments and in the Year 15 pace bridge sample.

Chapter 10.4 - Trend data

This section describes the extension of the IRT reading scale, defined originally on Year 15 BIB spiral data, to the paced data of the previous NAEP assessments.

Chapter 10.5 - Scale definition and behavioral anchoring

This section details the procedures by which results on the IRT reading scale were related to expected performance on specific reading tasks.

10.0.1 Item Response Theory

Item response theory (IRT) provides a mathematical model for the probability that a particular examinee will make a correct response to a particular item, in terms of a parameter reflecting the examinee's proficiency, and one or more parameters characterizing features of the item such as its difficulty and reliability (Lord, 1980). As an example, the three-parameter logistic IRT model (the model used in the NAEP reading assessment) takes the following form:

$$P(x_{ij}=1|\theta_i,a_j,b_j,c_j) = c_j + (1-c_j)/[1+\exp[-1.7a_j(\theta_i - b_j)]],$$

where

x_{ij} is the response of pupil i to item j, 1 if correct and 0 if incorrect,

θ_i is the (unobservable) proficiency of pupil i,

a_j is the slope parameter of item j, characterizing its sensitivity to proficiency,

b_j is its threshold parameter, characterizing its difficulty,

c_j is its lower asymptote parameter, reflecting possibly non-zero chances of correct response from even persons of very low proficiency, and

1.7 is a scaling constant.

The curve traced by this function for a given item as θ runs over its range is referred to as an "item response curve." A domain of items over which performance is modeled, and the accompanying proficiency variable, are often jointly referred to as a "scale."

IRT has effectively revolutionized measurement in education and psychology by virtue of the advantages it offers over traditional "true-score" or "number-right" test theory. Of particular note are (i) its capacity to provide comparable measurements from different item sets without expensive equating studies, (ii) its flexibility to administer examinees sets of items that are tailored to their proficiency levels, and (iii) its ability to yield scores that can be interpreted in terms of expected behavior on every item in the scale.

To date, applications of IRT have been limited for the most part to the setting of individual measurement. That is, each individual is administered a sufficient number of items to provide a precise estimate of his or her (unobservable) proficiency parameter, an estimate that is then used in subsequent decision-making or secondary analysis. It has been argued, however, that the advantages mentioned in the preceding paragraph hold promise for the assessment setting as well (Bock, Mislevy, & Woodson, 1982; Messick, Beaton, & Lord, 1983), despite the fact that interest lies not in the proficiencies of individual examinees but in proficiency distributions within targeted populations, the population trends over time, and their relationships with examinees' pedagogically and socially relevant background characteristics.

10.0.2 Item Response Theory and Educational Assessment

The source of interest in IRT for NAEP was dissatisfaction with the reporting methods that had evolved prior to the Year 15 assessment. When NAEP was conceived, the plan was to report for each individual item the estimated proportion of correct responses from a population or a subpopulation. This single-item reporting quickly proved too cumbersome, and by the second reading assessment NAEP reported averages of estimated percents-correct for sets of related items. Comparisons over time or across ages in terms of these "domain percents-correct" were necessarily limited to sets of items common to all groups involved in the comparison--a limitation strongly felt as the NAEP item pool evolved over time, thus reducing the number of items by which trends could be estimated. Interpretations of domain percents-correct were limited as well, since generalizations to different items sets or implications for particular items are not forthcoming. Finally, because different individuals were administered different items under NAEP's matrix-sampling design, nothing comparable to traditional test scores was obtained to facilitate secondary analysis of relationships among proficiency and other background variables.

Three objectives, then, were established for the use of IRT in NAEP:

(1) Results should be summarized in a manner which would facilitate comparisons over time and across subpopulations (including different ages and grades), despite the fact that

different item sets were administered to different targeted comparisons groups.

(2) Results should be reported on a scale that could be interpreted in terms of expected behavior on tasks involving reading.

(3) Secondary users should be provided results in a form that facilitates analyses of the relationships among reading proficiency and examinee characteristics, such as instructional experiences and demographic data.

The original intention was to accomplish these objectives by estimating each sampled student's proficiency variable on an IRT scale. Distributions of these estimates would be taken as approximations of the latent proficiencies themselves, both for NAEP reports and for secondary analyses. As documented in Chapter 10.2, however, this approach proved unsatisfactory, mainly because most pupils responded to too few cognitive exercise to provide precise point estimates of their latent proficiency variables. More complex methods that could provide estimates of population characteristics without estimating values for individual respondents had to be developed.

Anticipating and summarizing the contents of the chapter, the new methodologies accomplished objectives 1 and 2 in full. Objective 3, providing useful data for secondary analysts, is satisfied to a large but incomplete extent. The procedures that would be required to support all conceivable secondary analyses of NAEP data, to the level of accuracy inherent in the data, turn out to be beyond the reach of present (and indeed, foreseeable) resources. The procedures described in Chapter 10.3 do however possess the properties of (i) yielding consistent estimates on the IRT scale for results related to the traditional NAEP reporting variables, (ii) providing approximate, though sub-optimal, results for other background variables (potential biases of 15 to 40 percent in certain regression coefficients, for example), and (iii) laying the methodological foundation for improved estimation of background effects in future NAEP IRT analyses (reducing potential biases to a maximum of, say, 5 percent for a broad range of policy analyses).

Two points merit emphasis here. First, all analyses that could be carried with past NAEP data can still be carried out with equal or greater precision with the Year 15 data. Because item responses are provided on public-use data tapes, nothing is lost to the secondary analyst by the fact that some results are reported on an IRT scale.

Second, the biases mentioned in the preceding paragraph are not shortcomings of our procedures but of limitations inherent in the data, namely the sparseness of information about individual respondents. When it is desired to draw inferences from results on specific items to proficiencies of a more general nature, the biases of "errors in variables" problems arise. Typically, because they are difficult to deal with and are not well understood in the educational research community, they are ignored (as in analyses of the High School and Beyond). This standard practice would prove disastrous for trend analyses in NAEP data, since the resulting biases vary with the data-gathering design; the variation of NAEP's sampling design over time, due in part to varying levels of funding, would render useless any trend analyses that ignored these effects. The innovations described in Chapters 10.3 through 10.5, however, open the door to powerful and useful analyses based on IRT (e.g., The Reading Report Card, 1985), and in which errors in variables are handled appropriately.

IRT provides more powerful analyses than percent-correct reporting in large degree because it makes more assumptions about relationships among examinees' expected responses to items. The original justification for reporting single-item percents-correct, for example, was the fact that each item offers some unique information about trends and population comparisons. Nonetheless, trends or comparisons based on each of several items from the same content area will exhibit similarities--most geometry items might be becoming easier over time, for instance, while most algebra items are becoming more difficult. Fitting one unidimensional IRT model to algebra items and another to all geometry items will capture these commonalities, operationally defining the latent "algebra" and "geometry" proficiency variables in terms of the similarities of patterns of the items in a scale.

The cost of using the IRT models is the loss of information about differences among the patterns of items within a scale. If both algebra and geometry were modeled by a single scale in the example above, for instance, the IRT single-variable summary would not appropriately reflect the differential changes of items of the two types. Technically, model mis-specification errors of this type are referred to as "multidimensionality" or "lack of local independence." (See Goldstein, 1980, and Traub and Wolfe, 1981, for insightful discussions of the threat such errors pose to the use of IRT in educational assessment.) Similarly lost will be differential patterns of performance on the items within a scale for reasons of (i) differing curricula or teaching styles over schools, (ii) changes in curricular emphasis over time, and (iii) regional or ethnic-group differences.

This line of reasoning leads to three important conclusions. First, it is clear that summaries of assessment data in terms of IRT variables merely reflect the dominant patterns recurring within a much broader and richer data base. At best they serve as summary indicators like the Gross National Product or the Consumer Price Index; they will undoubtedly prove inadequate for more subtle analyses that demand differential information among items within scales, for comparing detailed effects of teaching methodologies or for analyzing item performance in terms of specific skill components demanded by particular exercises. (Witness, for example, Haertel's [1984] use of latent class models to study NAEP mathematics exercise in terms of the skills they demand.) IRT proficiency variables may be justified by their usefulness as a data reduction technique, but it must be borne in mind that they are not founded strictly in accordance with either pedagogical or psychological theories about the skills examinees bring to bear upon the exercises they are presented.

Second, because IRT variables are defined operationally within scales, pedagogical and psychological theories must play a role in determining the domains of items that will be scaled together. Because differential patterns within a domain will not be reflected by the IRT results, scaling should be carried out within domains for which broad summaries are sensible and policy-relevant. These decisions must be theory-driven as well as data-driven (see paragraph below). For this reason, "study skills" tasks requiring declarative knowledge were eliminated from the domain that became the basis of the NAEP reading scale. This focused the analysis on the more generalized skills commonly thought of as reading per se, among which different curricula or backgrounds were less likely to impose strong differential patterns of item performance.

Third and finally, the burden thus falls upon those who propose to use IRT in educational assessment to demonstrate (and to continue to demonstrate over time) that the domains of items within which they carry out IRT scaling are in fact capturing relevant patterns of change. This must be done by examining what are in a broad sense residuals from the IRT models: for example, factor analyses of items within scales, analyses of residuals from fitted item response curves, and examinations of the stability of item response curves over time. (Analyses of this type are described in Chapters 10.1 through 10.4.)

Chapter 10.1

ASSESSMENT OF THE DIMENSIONALITY OF NAEP YEAR 15 READING DATA[1]

Rebecca Zwick

Educational Testing Service

10.1.1 The Unidimensionality Assumption in Item Response Theory

To determine whether it was reasonable to regard the reading items administered in the Year 15 NAEP data collection as measures of a single construct, a series of analyses of the dimensionality of the reading data was performed. Dimensionality analyses were conducted both within and across the three grade/ages, 4/9, 8/13, and 11/17. It was important to investigate the dimensionality issue because the validity of the item response theory (IRT) model used to estimate reading proficiency in the 1983-1984 NAEP survey rests on the assumption of unidimensionality. It should be noted, however, that regardless of whether an IRT model is used, it is ordinarily assumed that items on an achievement test can be treated as measures of a single dimension, in this case, reading proficiency. Scoring a test by simply summing the item scores involves an implicit assumption of unidimensionality; IRT scaling formalizes this assumption.

The reading data were analyzed using the three-parameter logistic model (Birnbaum, 1968; Lord, 1980) in which P_{ij}, the probability that subject i gets item j correct can be expressed as follows:

$$P_{ij} \equiv P(x_{ij} = 1|\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta_i - b_j)}} \qquad [1]$$

where θ_i is the proficiency parameter for person i, a_j is the item discrimination parameter, b_j is the item difficulty, and c_j can be interpreted as the probability that a person with very low ability gets item j correct. (Model parameters were estimated using BILOG [Mislevy & Bock, 1982]; details are provided in Chapter 10.3.) In applying a model of this kind, it is assumed that the only examinee characteristic that affect item response is a single latent variable, θ.

---

[1] The author thanks Albert Beaton, Bruce Bloxom, Darrell Bock, Neil Dorans, Paul Holland, Robert Mislevy, Paul Rosenbaum, and Ledyard Tucker, who provided consultation and comments; Dick Harrison, Bruce Kaplan, and Dorothy Thayer, who programmed the analysis procedures; and Natalie Roca, who conducted analyses and literature reviews. An earlier version of this chapter is available as ETS Research Report 86-4.

10.1.1.1 Robustness of IRT Estimation Procedures

In practice, the assumption of unidimensionality, required for the application of conventional IRT models, will always be violated to some degree. To make a more objective determination as to what constitutes an important departure from unidimensionality, we need to know more about the robustness of the IRT estimation procedures to violations of the unidimensionality assumption. Although there has been little theoretical work in this area, some empirical studies have been conducted. Reckase (1979) and Drasgow and Parsons (1983) investigated the results of estimating the three-parameter logistic model, using LOGIST (M. S. Wingersky, 1983) under violations of the unidimensionality assumption. (The one-parameter and two-parameter logistic models were also examined by Reckase, 1979, and Drasgow and Parsons, 1983, respectively.) Reckase's study was based on five actual data sets and five data sets constructed to have specific factor structures. He concluded that LOGIST estimates "the first principal component when it is large relative to other factors .... good ability estimates can be obtained ... even when the first factor accounts for less than 10 percent of the test variance, although item calibration results will be unstable. For acceptable calibration, the first factor should account for at least 20 percent of the test variance" (p. 228). Drasgow and Parsons (1983) made use of a hierarchical model with a general latent trait as well as five group factors to simulate various kinds of latent structures. One of their conclusions was that, in the simulated data designed to resemble "moderately heterogeneous achievement tests and attitude assessment instruments" (p. 193), LOGIST still recovered the latent trait and provided acceptable estimates of the item parameters (p. 198). There is no reason to believe that the effects of multidimensionality on BILOG (Mislevy & Bock, 1982), which was used to scale the NAEP data, would differ from the results obtained with LOGIST (Mislevy, personal communication, October 1985). These findings suggest that IRT scaling procedures can produce satisfactory results under moderate departures from unidimensionality.

10.1.2. Methods of Dimensionality Assessment for Dichotomous Data

The traditional psychometric approach to the assessment of dimensionality is through factor-analytic methods. Factor analysis often produces satisfactory results when each of the variables is the score on a multi-item test. When each of the measures is the response to a dichotomously scored item, however, it is now well known that linear factor analysis of Pearson (phi) correlations does not, in general, yield a correct representation of the dimensionality of the item pool (see, e.g., Carroll, 1945, 1983; Hulin, Drasgow, & Parsons, 1983; McDonald & Ahlawat, 1974; Mislevy, 1986c). The fundamental problem is that in computing phi correlations, item responses are treated as true dichotomies. In applying a linear factor analysis model, we are hypothesizing that dichotomous variables are linear combinations of continuous latent variables with infinite range, a mathematical impossibility. In fact, the regression of a dichotomous item on a continuous latent variable must be nonlinear. The best linear approximation to the nonlinear regression will depend on the region in which the data are most dense (Mislevy, 1986c); that is, it will be related to the item mean, or difficulty (as defined in classical test theory). From this perspective, it is not surprising that linear factor analysis of dichotomous items often produces a second factor, typically called a difficulty factor, that is related to item difficulty, but appears to be unrelated to any substantive property of the items. There can, in fact, be more than one such spurious factor (as is the case for items that form a perfect Guttman scale), but ordinarily, only one is substantial in size.

As an alternative to phi coefficients, tetrachoric correlations between items can be obtained. In computing tetrachorics, it is assumed that the item responses are functions of underlying continuous variables that have a bivariate normal distribution. The model dictates that, for each item, individuals who have values greater than a certain threshold on the underlying response variable get that item correct; individuals with values lower than the threshold get it wrong. Using the bivariate normality assumption, the correlation between the unobserved continuous variables can be inferred from the 2 x 2 table of item responses. Tetrachoric correlations do not provide a valid measure of association if bivariate normality does not hold. Furthermore, the occurrence of guessing violates the above model, which postulates that the probability that an individual gets an item right is a function only of his value on the underlying response variable. When guessing does occur, factor analysis of tetrachorics can produce spurious factors (see Carroll, 1945, 1983; Hulin, Drasgow, and Parsons, 1983). Adjustments for guessing are theoretically possible, but often lead to unacceptable results in practice. (Attempts to adjust for the effects of guessing in the NAEP analyses are discussed in Section 10.1.3.2.1.) Additional problems are inaccuracies in the computation of tetrachorics as their absolute values approach unity, the large standard errors of the coefficients, and the occurrence of non-Gramian matrices of sample tetrachorics, even when data are complete. (In the case of the NAEP analyses, in which a large proportion of data are missing by design, the negative eigenvalues tend to comprise a large proportion of the trace of the tetrachoric matrix; see Section 10.1.3.1.2 and Table 10.1(3).)

It is clear that conventional factor analysis of phi and tetrachoric correlations is not a satisfactory means of investigating dimensionality. Unfortunately, no uniformly accepted statistical procedures for dimensionality assessment exist for the case of dichotomous variables. As a result, a vast literature on the subject has developed, particularly during the last ten years, as the use of IRT models has increased. Some methods which have gained attention recently are briefly described here; more detailed reviews of dimensionality assessment are given by Hattie (1984, 1985), Hulin, Drasgow, and Parsons (1983, Chapter 8), and Mislevy (1986c).

Factor-analytic methods that have been proposed to overcome the problems described above include factor analysis of item parcels, nonlinear factor analysis, the generalized least squares methods developed by Christofferson (1975) and Muthén (1978), and the full-information maximum likelihood method of Bock and his associates (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1985).

Factor analysis of item parcels is achieved by grouping items into meaningful subsets (the so-called parcels) and then applying conventional factor-analytic methods to the parcel scores. This method was applied by Cook and Eignor (1984) to a portion of the NAEP data collected in 1979-1980 and by Cook, Eignor, Dorans, and Petersen (1985) to SAT data. One practical problem with this approach is that it may be difficult to classify certain items a priori. Furthermore, if the item parcels differ in average difficulty, the obtained factor structure may be influenced to an undesirable degree by item difficulty, as in the dichotomous case (Kingston & Dorans, 1982). A more fundamental drawback is that this approach does not assess directly the properties of individual items. Because item scores do not enter the analysis, it is possible for items that measure a property other than the one of interest to go undetected. Finally, the application of this approach to the complete NAEP data set is virtually ruled out because examinees do not all receive the same items (see Section 10.1.3.1). (The Cook and Eignor [1984] analysis was based on a subset of examinees who had been administered the same items.)

In a series of publications, McDonald presented a theory of nonlinear factor analysis (e.g., McDonald, 1967, 1983). In McDonald's model, $P(x_{ij} = 1 \mid \underset{\sim}{\theta})$, the conditional probability that an examinee answers an item correctly, given his observed vector of latent traits, $\underset{\sim}{\theta}$, is expressed as a nonlinear function of the latent traits. For example, in one version of the model, $P(x_{ij} = 1 \mid \underset{\sim}{\theta})$ is expressed as a weighted sum of polynomial functions of the latent traits. Simulation studies of the effectiveness of nonlinear factor analysis as a method of dimensionality assessment have led to inconsistent findings. Hambleton and Rovinelli (1986) found that a one-factor polynomial model with linear and quadratic terms provided a good fit to a simulated unidimensional data set, unlike a one-factor linear model. Furthermore, a two-factor polynomial model provided a good fit to two-dimensional simulated data. Based on this and other findings, Hambleton and Rovinelli concluded that nonlinear factor analysis is one of the most promising methods for assessing the dimensionality of dichotomous data. On the other hand, Hattie (1984) concluded that the sum of absolute residual covariances from nonlinear factor analysis could not be recommended as an index of dimensionality because results from the unidimensional and multidimensional data sets were not sufficiently distinct.

Christofferson (1975) developed a factor-analytic method for dichotomous data that involves expressing the expected proportion correct for each item and for the joint proportions correct for each pair of items as a function of item thresholds (see above and Section 10.1.3.4, below) and factor loadings. The weighted distance between the observed and modeled values of these proportions is then minimized using generalized least squares (GLS) methods. Christofferson's solution makes use of the information contained in the three-and four-way margins of the n-way contingency table of item responses (see Appendix 2 in Christofferson, 1975; Mislevy, 1986c), unlike conventional factor analysis of phi or tetrachoric correlations, which makes use of only the one-and two-way marginals. Solving for estimates of the thresholds and loadings requires numerical integration and is therefore computationally burdensome. Muthén (1978) developed an alternative GLS method that reduces the computational requirements to some degree. However, application of both Christofferson's and Muthén's methods is currently limited to about 25 items. Bock and associates developed a factor-analytic approach for dichotomous data, called full-information factor analysis (Bock, Gibbons, & Muraki, 1985) because it uses information contained in the joint frequencies of all orders of the item responses. This method, detailed in Section 10.1.3.4 below, makes use of the marginal maximum likelihood methods of Bock and Aitkin (1981) for estimating the parameters of the common factor model.

In addition to factor-analytic approaches, a number of other methods of dimensionality assessment have been proposed. For example, Bejar (1980) has recommended comparing the estimated item difficulties (i.e., the estimates of the b_j of equation 1) obtained by calibrating a complete set of test items to those obtained by performing the calibration separately within content areas. (Bejar [1980] also proposed an additional procedure, which involves computing, for each content area, a scaled score corresponding to each of the two sets of item parameter estimates, and then comparing the results obtained by fitting a one-factor model to each of the two sets of scores.) Although Bejar's (1980) application of the method appeared to yield useful results, Hambleton and Rovinelli (1986) found that the method was unable to discriminate between one-and two-dimensional simulated data sets. Another method that has been proposed is analysis of the residual differences between the observed proportions of correct responses for individuals within various categories of proficiency and the estimated probabilities of correct responses according to the unidimensional item response model deemed appropriate (e.g., Equation 1). Various methods of residual analysis have been proposed; reviews are given by Traub and Wolfe (1981) and Hattie (1985). The rationale is that if the model fits well, the data can be assumed to be consistent with unidimensionality. A major drawback is that large residuals may be the result of model violations other than multidimensionality. Hambleton and Rovinelli (1986) concluded that indices based on the size of average residuals obtained after fitting one-, two-, and three-parameter logistic models were not capable of detecting multidimensionality. It should be noted that Hambleton and Rovinelli did not report any investigation of the pattern of residuals.

10.1.3 Methods Used to Assess the Dimensionality of NAEP Reading Data

The proposed methods of dimensionality assessment differ in terms of the assumptions needed, the hypothesis tested, and the statistical artifacts that affect interpretation. Rather than selecting a single method of dimensionality assessment for the NAEP reading data, we applied four different techniques, described in this section. For descriptive purposes, we included principal components analysis (PCA) of phi and tetrachoric correlations, as described in Section 10.1.3.2. As an experimental analysis, we also applied PCA to the image correlation matrix, a method based on the work of Guttman (1953) and Kaiser and Cerny (1979), described in Section 10.1.3.3. Bock's full-information factor analysis, discussed in Section 10.1.3.4, was applied to a subset of the data. Finally, we used the method of Rosenbaum (1984a, 1984b), described in Section 10.1.3.5, which involves examination of the partial association for each pair of items, conditional on the total score on the remaining items. Prior to a discussion of these methods, the properties of the NAEP database are described.

10.1.3.1 Properties of NAEP Data

10.1.3.1.1 Items Included in Dimensionality Analyses

All reading items that were included in the IRT scaling and were also spiralled with other items (see Section 10.1.3.1.2 and Chapter 10.2) were used in the dimensionality analyses. All subjects who responded to one or more of these items were included. The number of subjects and items available for the analyses is shown in Table 10.1(1). (The NAEP item numbers for all items included in the dimensionality analyses are given in Appendix 1 of this chapter.) As indicated, there were about 100 items per grade/age. Twenty-five of the items included in the analyses were administered to all three grade/ages. The range and mean of the proportions correct for each of the three grade/ages and for the 25 across-grade/age items are given in Table 10.1(1). As shown, the number of students per grade/age was roughly 26 to 29 thousand. As a result of the number of items and subjects in the data base, certain analyses were ruled out because they were too costly or exceeded computing capabilities. In other cases, dimensionality analyses were performed on only a subset of items to minimize the cost and the computational burden.

Ninety-four percent of the NAEP reading items included in the analyses were multiple choice items with three to six response choices. The remainder were essay items in which the respondent was asked to react to a reading passage. Essay items were scored on a scale of 1 to 5, which was later dichotomized. All items were classified by reading experts on the basis of objective (deriving information vs. integrating and applying information), stimulus (short or long reading passage, document, or picture), and content (fictional story, poem, informational passage, social studies, science, arts and humanities, or life skills). These item properties, as well as a further classification of the items based on the work of Mosenthal (1985), were used in attempting to interpret analysis results. (A subset of reading items that were designed to assess study skills were not included in the dimensionality analysis because they were not scaled using IRT. That these items differed from the remaining reading items was suggested by examination of the item content, as well as empirical evidence: For a subset of examinees, number-right scores on blocks of study skills items and on blocks of conventional reading items were obtained. The attenuation-corrected correlations between study skills blocks and conventional reading blocks tended to be lower than intercorrelations between conventional reading blocks. Many of the items which led to departures from unidimensionality in Jungeblut's [1984] analyses of the 1979-1980 NAEP data were study skills items [Jungeblut, personal communication, October 1985].)

10.1.3.1.2 Missing Data Pattern

A new feature of the Year 15 NAEP design was the use of balanced incomplete block (BIB) spiralling to assign test items to booklets (see Messick, Beaton, & Lord, 1983; Beaton, 1984; and Chapter 5). BIB spiralling combines the features of conventional spiralling and multiple matrix sampling. As in ordinary multiple matrix sampling, each item is administered a prescribed number of times, although examinees receive different subsets of items. BIB spiralling has the additional feature that each pair of items is assessed a prescribed number of times. In NAEP, reading items were first grouped into blocks, consisting in most cases of 8 to 12 items, which were then assigned to test booklets according to a design that provided the desired links between items. This resulted in a set of approximately 60 different test booklets per grade/age, which were assigned to respondents in a random sequence.

A major advantage of BIB spiralling is that it permits the estimation of inter-item correlations. However, the resulting matrix of correlations, referred to here as the BIB matrix, has an unusual pattern of missing data. In the case of the NAEP reading data, the number of respondents available to estimate correlations between items in the same block is, in most cases, nine times the number of respondents available for the estimation of correlations between items that fall within different blocks. Furthermore, the correlations of items in one block, say, A, with those in another block, B, are not in general based on the same group of respondents as the correlations of Block C items with Block D items. Because of the spiralling procedure used to assign booklets to respondents, the missing data that result from the implementation of a BIB design can be regarded as random. However, in using a BIB correlation matrix rather than a conventional correlation matrix, we are implicitly making the assumption that the correlations between items are not subject to context effects. If, for example, the population correlation between two items, i and j, varied depending on whether k were administered with i and j, then the sample correlation of i and j in the presence of k would not be an estimate of the same population parameter as the sample correlation of i and j in the absence of k. Computation of a BIB matrix involves averaging these sample correlations, which would be undesirable under these circumstances.

Even if the assumption of no context effects is justified, there are other ways in which the properties of the BIB matrix differ from those of a conventional correlation matrix. For example, the standard errors of the within-block correlations are smaller than those of the between-block

Table 10.1(1)

Number of Items and Students Available for Dimensionality Analyses

Grade/Age	Number of Items	Proportions Correct Minimum	Proportions Correct Maximum	Proportions Correct Mean	Number of Students
4/9	108	.04	.93	.50	26,087
8/13	100	.09	.98	.63	28,405
11/17	95	.21	.96	.70	28,861
Across Grade/Ages (Common Items)	25	.13	.90	.53	83,353

correlations. Also, the BIB matrix may have negative eigenvalues, unlike a conventional correlation matrix. As detailed in Section 10.1.3.1 and Tables 10.1(2) and (3), both phi and tetrachoric matrices of NAEP items had negative roots in most cases. For analyses that required a matrix that was at least positive semi-definite, an adjustment procedure, described in Appendix 2 of this chapter, was applied. Although there is no indication that analysis results were affected in any major way by the use of BIB matrices or their adjusted counterparts, the statistical properties of these matrices are not fully understood at present. The special properties of BIB matrices and the impact of BIB spiralling on the NAEP dimensionality analyses are discussed in further detail in Section 10.1.4.

In addition to the BIB missing data, which can be regarded as random, there are two major categories of non-random missing data: omitted items and items that the respondent was administered but did not reach. Unanswered items occurring after the last valid response within a block were considered "not reached." (In administering the items, each block was timed separately.) Unanswered items that occurred prior to the last valid response (and were not a result of the BIB design) were coded as omits. The category of omitted items was defined to include as well any items marked, "I don't know," which was a response alternative for all multiple choice items. The treatment of not reached and omitted items in each of the dimensionality analyses is discussed in Sections 10.1.3.2 to 10.1.3.5.

10.1.3.2 Principal Component Analysis of Inter-item Correlation Matrices

Despite the drawbacks described in Section 10.1.2, principal component analyses (PCA) of the phi and tetrachoric matrices for each grade/age were conducted for descriptive purposes. In addition, analyses including all respondents were performed, based on the 25 items common to all three grade/ages. It can be argued that the results of these analyses represent a "worst case"; that is, because the analyses tend to produce spurious factors, results that were free of artifacts would be expected to be more consistent with unidimensionality.

Items that were not reached were excluded from the analysis; omitted items were scored as incorrect. For each of the four phi matrices, Table 10.1(2) gives the range of inter-item correlations, the median correlation, the first five eigenvalues and the percent of the trace they represent, and, as an index of the degree to which the matrix departed from positive-definiteness, the sum of the negative eigenvalues as a percent of the trace of the matrix. The median sample size (N) on which the correlation coefficients were based (see Section 10.1.3.1.2) is also given. The corresponding information for the tetrachoric matrices is given in Table 10.1(3). The results in Tables 10.1(2) and (3) are based on analyses that incorporated the respondents' sampling weights (see Chapter 13.1). Unweighted analyses yielded almost identical results.

It is clear that, for each of the eight matrices, there is a large first root, constituting between 17 and 25 percent of the trace for the phi matrices and between 30 and 40 percent for the tetrachoric matrices (but

Table 10.1(2)

Eigenvalues and Descriptive Statistics for Phi Matrices

Grade 4/Age 9 (108 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
23.9	22	Median N	280
3.3	3		
2.5	2	Range of r	-.18, .53
2.4	2	Median r	.19
2.2	2	Neg. roots as pct. of trace	3

Grade 8/Age 13 (100 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
17.0	17	Median N	323
2.6	3		
2.5	2	Range of r	-.15, .60
2.2	2	Median r	.14
2.1	2	Neg. roots as pct. of trace	2

Grade 11/Age 17 (95 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
17.5	18	Median N	331
3.1	3		
2.3	2	Range of r	-.16, .68
2.1	2	Median r	.16
2.0	2	Neg. roots as pct. of trace	2

All Grade/Ages Combined (25 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
6.3	25	Median N	919
1.5	6		
1.2	5	Range of r	.29, .57
1.1	5	Median r	.18
1.0	4	Neg. roots as pct. of trace	0

Table 10.1(3)

Eigenvalues and Descriptive Statistics for Tetrachoric Matrices

Grade 4/Age 9 (108 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
39.5	37	Median N	280
6.6	6		
4.7	4	Range of r	-.46, .81
3.7	3	Median r	.35
3.4	3	Neg. roots as pct. of trace	27

Grade 8/Age 13 (100 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
30.0	30	Median N	323
4.3	4		
3.8	4	Range of r	-.34, .81
3.4	3	Median r	.27
3.3	3	Neg. roots as pct. of trace	21

Grade 11/Age 17 (95 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
32.0	34	Median N	331
3.9	4		
3.3	3	Range of r	-.38, 90
3.0	3	Median r	.31
2.8	3	Neg. roots as pct. of trace	19

All Grade/Ages Combined (25 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
10.0	40	Median N	919
1.6	6		
1.2	5	Range of r	.05, .80
1.2	5	Median r	.33
1.0	4	Neg. roots as pct. of trace	0

note that the negative roots constitute up to 27 percent of the trace for tetrachoric matrices). The second root is always less than one-fourth of the first. Following the sharp drop-off between the first and the second, the remaining roots trail off gradually. These findings are reassuring in that they are consistent with a large first dimension. (The size of the first component may appear small to those who are unaccustomed to examining the results of item-level factor analyses. In interpreting these findings, however, it is important to consider that the median inter-item correlations are low: between .14 and .19 for the four phi matrices and between .27 and .35 for the tetrachoric matrices. Results of PCA of phi matrices computed from simulated unidimensional data showed that the first root typically constituted 25 to 30 percent of the trace; see Section 10.1.3.3 and Table 10.1(5).) The loadings on the first principal component were not related in any obvious way to the item classifications discussed in Section 10.1.3.1.1.

10.1.3.2.1 Application of Guessing Corrections to Tetrachoric Correlations

When it is possible for items to be answered correctly through guessing, the magnitude of observed tetrachoric correlations is related to item difficulty (e.g., see Hulin, Drasgow, & Parsons, 1983, pp. 249-255). To eliminate this problem, Carroll (1945) suggested that the frequencies in the 2 x 2 tables of responses for each pair of items be adjusted to "remove" the effects of guessing and that tetrachorics be computed on the basis of these adjusted frequencies. In Carroll's model, it is implicitly assumed that, for each pair of items, the probability of getting one item right by guessing is independent of the probability of making a correct guess on the other item. In applying the model, it is typically assumed that guessing is random and that the probability of getting an item right by guessing is therefore equal to the reciprocal of the number of response choices. To determine whether it would be a useful strategy for NAEP data, Carroll's correction was applied to the item responses for Grade 8/Age 13, setting g_j, the hypothetical probability of guessing right on item j, equal to the reciprocal of the number of response choices for item j, excluding the "I don't know" alternative. For essay items, g_j was set to 0. The results were clearly unsatisfactory: It was found that 16 percent of the tetrachoric coefficients were rendered incomputable because of negative adjusted cell frequencies. Several other corrections were investigated, but deemed unsatisfactory, including a modification of Carroll's correction in which the input g_j values were adjusted so as to avoid the occurrence of negative adjusted cell frequencies and a correction in which each g_j was set equal to the estimated lower asymptote, c_j (see equation 1) of the item from the IRT item calibration. (Note that Bock, Gibbons, & Muraki [1985] describe a modification of Carroll's correction that apparently produces satisfactory results. This modified correction did not come to our attention until after our analyses were complete.)

10.1.3.3 Principal Components Analysis of the Image Correlation Matrix

Guttman (1953) developed a theory for the structure of quantitative variates called image theory. Image theory is based on the partitioning of a variable into two additive segments: the part that can be predicted through least squares linear regression of that variable on all the remaining variables, called the image, and the error of prediction, called the anti-image. Thus, unlike common factor theory, image theory provides an explicit definition for the common part of a variable. Another difference from the traditional factor-analytic approach is that, in general, the anti-images have non-zero covariances. Guttman shows that common factor theory may be viewed as a special case of image theory. The relation between image theory and other factor-analytic approaches is further examined by Harris (1962) and reviewed by Mulaik (1972).

Suppose that n variables are to be observed. The decomposition of the original variates into images and anti-images can be expressed as

$$\underset{\sim}{z} = \underset{\sim}{v} + \underset{\sim}{u} \qquad [2]$$

where $\underset{\sim}{z}$ is the n x 1 vector of observable random variables, standardized to have mean zero and unit variance, $\underset{\sim}{v}$ is the n x 1 vector random variable of images defined in equation 3, below, and $\underset{\sim}{u}$ is the n x 1 vector random variable of anti-images, or errors of prediction. (When referring to a finite sample of variables, Guttman used the terms partial image and partial anti-image. The qualifier, "partial" will not be used here.) The n x 1 vector random variable $\underset{\sim}{v}$ of images can be expressed as

$$\underset{\sim}{v} = \underset{\sim}{W}\underset{\sim}{z} \qquad [3]$$

The weight matrix $\underset{\sim}{W}$ is defined as

$$\underset{\sim}{W} = \underset{\sim}{I} - \underset{\sim}{S}^2\ \underset{\sim}{R}^{-1} \qquad [4]$$

where $\underset{\sim}{R}$ is the correlation matrix of the original variates, z, and

$$\underset{\sim}{S}^2 = [\mathrm{diag}\ (\underset{\sim}{R}^{-1})]^{-1} \qquad [5]$$

The off-diagonals of $\underset{\sim}{W}$ contain the regression weights for predicting each of the variates z from the remaining n - 1 variates. The diagonals of $\underset{\sim}{W}$ are equal to zero because the regression of a variate on itself is not of interest.

The principles of image theory are usually applied in practice by factor-analyzing $\underset{\sim}{G}$, the covariance matrix of the images, given by

$$\begin{aligned} \mathbf{G} &= \underline{E}(\mathbf{v}\mathbf{v}') = \underline{E}(\mathbf{W}\mathbf{z})(\mathbf{W}\mathbf{z})' \qquad [6] \\ &= \underline{E}(\mathbf{W}\mathbf{z}\mathbf{z}'\mathbf{W}') = \mathbf{W}\ \underline{E}(\mathbf{z}\mathbf{z}')\ \mathbf{W}' \\ &= \mathbf{W}\mathbf{R}\mathbf{W}' = (\mathbf{I} - \mathbf{S}^2\mathbf{R}^{-1})\ \mathbf{R}\ (\mathbf{I} - \mathbf{S}^2\ \mathbf{R}^{-1})' \\ &= \mathbf{R} + \mathbf{S}^2\ \mathbf{R}^{-1}\ \mathbf{S}^2 - 2\mathbf{S}^2 \end{aligned}$$

The j^{th} diagonal element of this matrix is the variance of the j^{th} image, which is equal to the squared multiple correlation coefficient (SMC) obtained by regressing the j^{th} variate on the remaining n - 1 variates. In this sense, $\mathbf{G}$ resembles the "reduced correlation matrix" of common factor analysis with SMCs used as communality estimates. The off-diagonals of $\mathbf{G}$, however, tend to be slightly smaller than those of the reduced correlation matrix (Kaiser, 1963); furthermore, $\mathbf{G}$ is always Gramian (assuming data are complete), unlike a correlation matrix with SMCs inserted in the diagonal.

As an alternative to the analysis of the $\mathbf{G}$ matrix, Kaiser and Cerny (1979) recommended principal component analysis of the image correlation matrix, $\mathbf{G}^*$, given by

$$\mathbf{G}^* = \mathbf{D}^{-1/2}\ \mathbf{G}\ \mathbf{D}^{-1/2} \qquad [7]$$

where

$$\mathbf{D} = \text{diag}\ (\mathbf{G}) = \mathbf{I} - \mathbf{S}^2 \qquad [8]$$

Kaiser (1970; see also Kaiser & Cerny, 1979) conjectured that image analysis would be well-suited to the factor analysis of dichotomous data. He noted that because the images are least squares predicted values of one variate based on the remaining n - 1 variates, "a crude appeal to the Central Limit Theorem suggests that the images will be sensibly multivariate normal, a set-up which is well known not to produce difficulty factors" (Kaiser, 1970, p. 407).

As an experimental approach to dimensionality assessment, principal component analysis of the image correlation matrix was applied to the NAEP data for Grade 4/Age 9, Grade 8/Age 13, and Grade 11/Age 17 and to the 25 across-grade/age items. Modification of the standard equations of image analysis was required because, in the case of NAEP data, the matrix $\mathbf{R}$ of weighted phi correlations is not positive definite (see 10.1.3.2 and Table 10.1(2)) and therefore can not be inverted. An adjustment procedure, detailed in Appendix 2 of this chapter, was used to obtain a singular approximation to the matrix of inter-item correlations and a pseudo-inverse of this adjusted matrix. Following this, the pseudo-inverse matrix $\mathbf{R}^-$ was then substituted for $\mathbf{R}^{-1}$ in the formulas for $\mathbf{W}$ and $\mathbf{S}^2$ (equations 3 and 4), as recommended by Kaiser and Cerny (1978). Analogues of the matrices $\mathbf{G}$, $\mathbf{G}^*$, and $\mathbf{D}$ (equations 6, 7, and 8) were computed using these modified forms of $\mathbf{W}$ and $\mathbf{S}^2$.

Results of the image analysis were superficially appealing. As shown in Table 10.1(4), the first roots of the image correlation matrix were often considerably larger than those of the phi matrix. For example, they were almost three times as large in the across-grade analysis. However, as described below, both empirical and theoretical examinations of this method show that it cannot provide the correct answer about dimensionality in the dichotomous case.

To investigate the properties of the image analysis solution, PCA of the image correlation matrix was applied to several simulated data sets generated from a unidimensional model. The simulation studies were conducted as follows:

(1) Assuming a three-parameter logistic model, NAEP reading items were calibrated with the LOGIST program (M. S. Wingersky, 1983) using actual NAEP data. Thirty of these items were randomly selected for this simulation run.

(2) One thousand pseudo-random values from a normal distribution with mean zero and unit variance were then generated. These represent theta or proficiency values for N = 1000 examinees.

(3) The three-parameter logistic function (Equation 1) was used to obtain the $n \times N = 30 \times 1000$ values of P_{ij}, the probability that person i gets item j correct. The item parameters a_j, b_j, and c_j were obtained from step 1 and the θ_i values from step 2.

(4) Corresponding to each value of P_{ij}, a pseudo-random value U_{ij} was generated from a uniform distribution on the interval [0,1]. If U_{ij} was less than P_{ij}, item j was scored as correct for person i; otherwise it was scored as incorrect. The correlation matrix of these simulated data was then obtained and the image procedure applied.

Table 10.1(5) shows the first five roots of the phi and image correlation matrices for one of the simulated data sets. Whereas the first root of the phi matrix was only about one quarter of the trace in the simulation, the first root of the image correlation matrix was about 80 percent of the trace. Other simulated unidimensional data sets produced similar values. If the size of the first root is used as a criterion, the image analysis technique appears to be superior to PCA of the phi matrix in revealing the true unidimensional structure underlying the data. However, as in the case of the phi matrix, the loadings of items on the second principal component of the image correlation matrix have substantial correlations with the proportions correct for the items: the correlations were .85 for the phi matrix and .65 for the image correlation matrix. This makes it clear that the image approach does not eliminate the problem of difficulty factors.

Table 10.1(4)

Eigenvalues of the Image Correlation Matrix

Grade 4/Age 9 (108 items)

First 5 Roots	Pct. of trace
27.3	25
9.5	9
3.7	3
3.2	3
2.7	3

Grade 8/Age 13 (100 items)

First 5 Roots	Pct. of trace
23.2	23
9.5	9
3.9	4
2.8	3
2.6	3

Grade 11/Age 17 (95 items)

First 5 Roots	Pct. of trace
25.8	27
5.7	6
4.3	4
3.4	4
3.3	3

All Grade/Ages Combined (25 items)

First 5 Roots	Pct. of trace
18.0	72
2.0	8
1.1	5
0.7	3
0.6	2

Table 10.1(5)

First Five Eigenvalues of Correlation and Image Correlation Matrices for Simulation Data (30 Items with NAEP Item Parameters)

Phi Matrix		Image Correlation Matrix	
First 5 Roots	Pct. of Trace	First 5 Roots	Pct. of Trace
7.7	26	23.8	79
1.7	6	2.6	9
1.1	4	0.5	2
1.0	3	0.5	2
1.0	3	0.4	1

Correlation of Loadings on Second Principal Component with Proportions Correct

.85	.65

Upon consideration, it seems unrealistic to expect the image approach to produce an accurate reflection of the number of dimensions, when it is known that factoring the phi matrix does not. After all, the image covariance matrix $\underset{\sim}{G}$ can be expressed as the sum of three terms, each of which is a function of the phi matrix. In addition, application of the image approach to dichotomous data involves the assumption of a linear regression model which is known to be violated. McDonald and Ahlawat (1974) expressed doubts about the use of image analysis in the dichotomous case, noting the relations between the eigenvalues of $\underset{\sim}{G}$ and those of $\underset{\sim}{R} - \underset{\sim}{S}^2$, the reduced correlation matrix with SMCs as communality estimates (see Harris, 1962).

Because it was evident from both a theoretical and an empirical perspective that the image approach produces misleading results in the dichotomous case, attempts to interpret the findings were discontinued.

10.1.3.4 Full-information Factor Analysis

A factor-analytic method that was designed for dichotomous data is full-information factor analysis (Bock, Gibbons, & Muraki, 1985; see also Mislevy, 1986c), which is implemented in the TESTFACT program (Wilson, Wood, & Gibbons, 1983). Unlike the methods described in Sections 10.1.3.2 and 10.1.3.3, this method does not require the computation of correlation coefficients, but operates instead on the set of distinct item response vectors. In contrast to factor analysis of correlation coefficients, which makes use of only the pairwise joint frequencies of item responses, Bock's full-information solution uses information contained in the joint frequencies of all orders. In applying this method, a particular model for the item responses must be assumed. In the case of the NAEP data, the selected model was a multivariate generalization of the three-parameter normal ogive in which each item is allowed to load on multiple factors. The model can be developed by first assuming that underlying the response of person i to item j is a response process variable defined as

$$y_{ij} = \sum_{k=1}^{K} \lambda_{jk} \theta_{ki} + \nu_j \qquad [9]$$

where θ_{ki} represents the value of the k^{th} latent variable (factor), k = 1, 2, ... K, for the i^{th} individual, i = 1, 2, ... N, λ_{jk} is the loading of the j^{th} item, j = 1, 2, ... n, on the k^{th} latent variable, and ν_j is a residual term associated with item j. The response process variables are assumed to have mean zero and variance one. The observed score of the i^{th} examinee on the j^{th} item, x_{ij}, takes on a value of 1, indicating a correct score, if y_{ij} exceeds γ_j, the threshold for the j^{th} item. Otherwise, $x_{ij} = 0$. If it is assumed that the residuals ν_j are independently distributed as $N(0,\sigma_j)$, the conditional probability that the i^{th} examinee gets the j^{th} item correct, given that his values on the latent variable are equal to $\underset{\sim}{\theta}_i$ can be expressed as

$$P(x_{ij} = 1 \mid \underset{\sim}{\theta}_i) = \frac{1}{\sqrt{2\Pi}\,\sigma_j}\int_{\gamma_j}^{\infty} \exp\left[-1/2\left(\frac{y - \sum\limits_{k=1}^{K} \lambda_{jk}\theta_{ki}}{\sigma_j}\right)^2\right] dv \qquad [10]$$

$$\equiv F_j(\underset{\sim}{\theta}_i)$$

This is a multivariate generalization of the two-parameter normal ogive model (see Lord & Novick, 1968).

This model can be modified to allow for the possibility of guessing by substituting

$$F_j^*(\underset{\sim}{\theta}_i) = c_j + (1 - c_j)\, F_j(\underset{\sim}{\theta}_i) \qquad [11]$$

for $F_j(\underset{\sim}{\theta}_i)$, where c_j represents the probability that an individual with very low ability gets the item correct. This multivariate generalization of the three-parameter normal ogive model was applied in the NAEP analyses. The c_j values are treated as fixed constants in the full-information factor analysis. The c_j parameters were estimated a priori using BILOG (Mislevy & Bock, 1982) and then input to the TESTFACT program. NAEP items that were coded as "not reached" (see Section 10.1.3.1.2) were not included in the analysis. Omitted items, on the other hand, were scored correct with probability c_j. Under this strategy, examinees who omit an item have the same theoretical probability of getting the item correct as examinees who guess in the absence of any information.

Incorporating the item response function, $F_j^*(\underset{\sim}{\theta}_i)$, defined in Equation 11, the marginal probability of the s^{th} response pattern can be expressed as:

$$P_s = P(\underset{\sim}{x} = \underset{\sim}{x}_s) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \prod_{j=1}^{n} F_j^*(\underset{\sim}{\theta})^{x_{sj}}\,[1 - F_j^*(\underset{\sim}{\theta})]^{1-x_{sj}}\, f(\underset{\sim}{\theta})\,d\underset{\sim}{\theta} \qquad [12]$$

where x_{sj} is the response to the j^{th} item in the s^{th} response pattern, $s = 1, 2, \ldots S$, and $S \leq \min(2^n, N)$ is the number of response patterns. It is further assumed in this application that $f(\underset{\sim}{\theta})$is the multivariate normal distribution with mean $\underset{\sim}{0}$ and covariance matrix $\underset{\sim}{I}$. Now, if it is assumed that the counts of the distinct response patterns follow a multinomial distribution, the likelihood of the matrix $\underset{\sim}{X}$ of observed counts r_s of distinct response patterns can be expressed as:

$$P(\underset{\sim}{X}) = \frac{N!}{r_1!\,r_2!\,\ldots\,r_s!}\; P_1^{r_1}\, P_2^{r_2} \ldots P_s^{r_s} \qquad [13]$$

where P_s is given by Equation 12.

The quantities P_s are approximated using numerical integration. The marginal maximum likelihood method of Bock and Aitkin (1981), which is based on earlier work by Bock and Lieberman (1970), is then applied to Equation 13 to obtain estimates of the factor loadings and thresholds for each item (see Bock, Gibbons, & Muraki, 1985; Mislevy, 1986c).

If sample size is sufficiently large, a test of the fit of the K-factor model relative to a general multinomial alternative can be obtained using a chi-square approximation to the likelihood ratio test. The model can be re-estimated and the test repeated for successive values of K. The difference between these chi-square statistics is also distributed as chi-square (under the hypothesis that the more restrictive model is correct) and can be used to test the improvement in model fit that is achieved by allowing the number of latent variables to increase. The test of change in model fit has been shown to perform well even when the frequency table is sparse (Haberman, 1977).

Because the TESTFACT program is very expensive to run, full-information factor analysis was applied only to 42 items for Grade 8/Age 13. These items, which were chosen to maximize the chances of detecting multidimensionality, were intended to represent four distinct item types: general reading comprehension, inference of word meaning from context, life skills, and essay. The comprehension, word meaning, and essay items all referred to passages the examinee was asked to read. Some passages were fictional stories; others pertained to an academic content area, such as science or social studies. The life skills items were based on documents that might be encountered in everyday life, such as a portion of a telephone directory, a grocery store coupon, or an advertisement. Responses to these 42 items were sent to Bock and his associate, Michele Zimowski, who conducted the analysis.

The analysis was based on the raw rather than the weighted frequency table of item responses. Because sampling weights have little effect on variances and covariances, they are unlikely to have much effect on factor analysis results (Bock, personal communication, November 1985).

Examination of the results led to the conclusion that a one-factor solution could be retained. The single factor accounted for about 39 percent of the total variance. In the unrotated two-factor solution, the first factor accounted for about 36 percent of the total variance; the second factor accounted for only 4 percent. Promax rotation (Hendrickson & White, 1964) resulted in a correlation of .77 between the factors. (The chi-square value for the improvement in fit achieved by adding a second factor was 78, with 41 degrees of freedom. If a design effect correction is incorporated [see Bock, Gibbons, & Muraki, 1985; Felleggi, 1979] based on the mean design effect for Grade 8 [see Chapter 14.2], the second factor narrowly misses statistical significance at $\alpha = .05$.) In the single factor solution, reading comprehension items, particularly those that involved fictional stories, tended to have the highest factor loadings. Life skills items had the lowest loadings.

10.1.3.5 Rosenbaum's Test of Unidimensionality, Monotonicity, and Conditional Independence

Rosenbaum (1984a) proves a theorem that states that if item characteristic curves are nondecreasing functions of a single latent variable, then conditional (local) independence of item responses, given the latent variable, implies certain relations among the item responses. Specifically, the conditional covariances between all monotone increasing functions of a set of item responses, given any function of the remaining item responses, will be non-negative. This theorem can be used to develop statistical tests of whether an observed data set is consistent with the assumptions of monotonicity, unidimensionality, and conditional independence. (See Holland, 1981, Holland & Rosenbaum, in press, and Stout, 1984, for further discussion of tests of this kind.)

As a special case of Rosenbaum's theorem, we can test the partial association for each pair of items, given number-right score on the remaining items, using the Mantel-Haenszel (1959) test, a conventional procedure for analysis of discrete data. In this case, we are examining the conditional covariance between monotone item summaries which are simply responses to a single item. The function on which we are conditioning is the number-right score on the remaining n - 2 items. To perform the Mantel-Haenszel test for a particular item pair, a 2 x 2 table of item responses is constructed for each of the K possible values of number-right score on the remaining items. Let n_{ijk} be the observed count in the i^{th} row, j^{th} column and k^{th} table, where $i = 1, 0$; $j = 1, 0$; and $k = 1, 2, \ldots K$. The Mantel-Haenszel test statistic is given by

$$Z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V(n_{11+})}}$$

where $E(n_{11+})$ and $V(n_{11+})$ denote the hypergeometric expectation and variance of $n(_{11+})$, given by

$$E(n_{11+}) = \sum_{k=1}^{K} \frac{n_{1+k}\, n_{+1k}}{n_{++k}} \qquad [15]$$

$$V(n_{11+}) = \sum_{k=1}^{K} \frac{n_{1+k}\, n_{0+k}\, n_{+1k} n_{+0k}}{n_{++k}^{2}\,(n_{++k} - 1)} \qquad [16]$$

and the plus subscript indicates summation over that subscript. The approximate significance level is obtained by referring Z to the lower tail of the standard normal distribution. A statistically significant result indicates that the pair of items has a negative partial association and is thus inconsistent with the hypothesized class of models.

The Mantel-Haenszel approach was programmed to accommodate the complexities of BIB spiralling in the following way: Suppose that we are interested in assessing the conditional covariance between items X_1 and X_2 and that, because of BIB spiralling, certain students who received items X_1 and X_2 also received X_3, X_4, and X_5, whereas others received X_5 and X_6. The test of association between X_1 and X_2 is then based on seven 2 x 2 tables: four corresponding to the possible score values for $X_3 + X_4 + X_5$ and three for the possible scores for $X_5 + X_6$. Because of the spiralling method used to assign booklets to respondents (see Section 10.1.3.1.2), the fact that respondents did not all receive the same items or even the same number of items does not impair the validity of the method. Items that were omitted or were administered but not reached (see Section 10.1.3.1.2) were scored as incorrect.

Because of the cost of computations, the Rosenbaum method was applied to only a subset of the NAEP items: those in blocks H, K, M, N, and O. The number of items per grade/age was 56 for Grade 4/Age 9, 53 for Grade 8/Age 13, and 56 for Grade 11/Age 17. The number of hypothesis tests, which is equal to the number of item pairs, was 1540, 1378, and 1540 for grade/ages 4/9, 8/13, and 11/17, respectively. To evaluate the findings of this method, a decision must be made about the appropriate alpha level at which to test these multiple hypotheses. Whereas on one hand, we would like to control the overall Type I error rate at an acceptable level, we do not want to maintain such rigorous Type I error control that a rejection of the hypothesis of unidimensionality would be impossible. As it turns out, even if the alpha for each hypothesis test is set at .01, a liberal alpha level for so large a number of tests, the number of statistically significant negative partial associations is only 4 for Grade 4/Age 9, 4 for Grade 8/Age 13, and 6 for Grade 11/Age 17. If alpha is set at .05 for each test, the number of statistically significant results is 31, 29, and 26 for the three grade/ages, respectively (see Table 10.1(6)). (It may at first seem surprising that less than 100α percent of the item pairs had statistically significant negative partial associations. However, note that we would expect to find 100α percent to be significant if all the conditional covariances were equal to zero in the population. If they are, in fact, greater than zero, less than 100α percent are expected to be significantly negative.) Therefore, it is reasonable to retain the hypothesis that the item responses can be represented by a monotonic unidimensional latent variable model with conditional independence. It should be noted that application of the Rosenbaum method does not provide a test of the fit of the three-parameter logistic model or of any other specific model.

In applying the Rosenbaum method, no modifications were incorporated to reflect NAEP's complex multi-stage cluster sampling scheme (Lago, Burke, Tepping, & Hansen, 1985). That is, raw rather than weighted frequencies were used in the analysis and no jackknifing or design effect adjustment was used in computing the significance probabilities of the Mantel-Haenszel statistics. As noted in Section 10.1.3.2, weighted and unweighted correlation matrices for the NAEP data are virtually identical, suggesting that the weights would make little difference in the Rosenbaum analyses.

Table 10.1(6)

Results of Rosenbaum Analyses

Within-Grade/Age Analyses

	Grade/Age		
	4/9	8/13	11/17
Number of items	56	53	56
Number of item pairs	1540	1378	1540
Number of significant negative partial associations:			
α = .01 per comparison	4	4	6
α = .05 per comparison	31	29	26

Across-Grade/Age Analyses

	Grade/Age Pair		
	4/9 & 8/13	4/9 & 11/17	8/13 & 11/17
Number of comparisons	24	24	24
Number of significant negative partial associations:			
α = .05 per comparison	0	0	0

Furthermore, the design effect for these tests is likely to be greater than one, as in 10.1.3.4. Adjustment of the significance tests would then lead to a reduction in the number of item pairs found to have negative partial associations, thus reinforcing the original conclusion about dimensionality.

10.1.3.5.1 Across-Grade/Age Analyses

In addition to determining whether it was reasonable to regard the reading items as unidimensional within each grade/age, it was of interest to investigate whether unidimensionality would hold if respondents from all three grade/ages were included. Of the entire set of items available for dimensionality analyses (Table 10.1(1)), 25 were administered to all three grade/ages. Twenty-four of these 25 were in the item blocks (H, K, M, N, O) used for the Rosenbaum analyses. A method developed by Rosenbaum (1984b), which is a variant of the approach described above, was applied to these 24 items. The procedure provides a test of whether the item responses of two groups of examinees is consistent with a difference in the distribution of a unidimensional latent variable. A rejection of this hypothesis may indicate the existence of additional dimensions. As a first step in the analysis, an indicator variable is created to represent group membership, with the higher value associated with the group hypothesized to have generally higher values on the latent variable. If the pattern of item responses is consistent with the hypothesized model, the conditional covariances of each item with the indicator variable will be non-negative, as described in 10.1.3.5.

For the NAEP data, a separate analysis was conducted for each pair of grade/ages, as follows: An indicator variable representing grade/age was created, with a value of 1 indicating the higher grade/age and the value of 0 corresponding to the lower grade/age. The partial association of each of the 24 items with grade/age was then assessed, using the Mantel-Haenszel (1959) test, as described in 10.1.3.5. With an alpha of .05 for each of the 24 hypothesis tests per grade/age pair (see Table 10.1(6)), no significant negative partial associations of items with the dummy-coded grade/age variable were found. This means that, as we would expect intuitively, students in higher grade/ages were more likely than students in lower grade/ages to answer items correctly, conditional on number-right score on the remaining items. These results are consistent with unidimensionality of the item pool.

10.1.4 The Impact of BIB Spiralling on Dimensionality Analyses

As noted above, the missing data that results from the BIB design can be regarded as random. However, this in itself does not imply that the results of NAEP data analyses are unaffected by BIB spiralling. In this section, the impact of BIB spiralling on each of the NAEP dimensionality analyses is considered.

The principal components analyses of the phi, tetrachoric, and image correlation matrices make use of BIB correlation matrices. These matrices have several properties that distinguish them from conventional correlation matrices. For example, the standard errors of the within-block correlations are smaller than those of the between-block correlations. Also, BIB matrices may have negative eigenvalues, unlike conventional Pearson correlation matrices.

To investigate the properties of BIB matrices, a series of simulation studies was conducted, one of which is reported here. Unidimensional item responses for 1,000 "subjects" on 30 items were generated using the procedures described in Section 10.1.3.3. The first 10 items were arbitrarily designated as block A, the second 10 as block B, and the third 10 as block C. Two correlation matrices were then computed. The first was an ordinary phi matrix, computed using the complete matrix of item responses. The second was computed by censoring the item responses according to a BIB design in which the first 333 examinees received blocks A and B, the next 333 received blocks B and C, and the remaining 334 examinees received blocks A and C. Pairwise correlations between all items were then computed. Table 10.1(7) shows which subjects were available to estimate the within-and across block correlations in the BIB matrix. For example, the correlations between items within block A were estimated using examinees 1-333 (who received blocks A and B) and 667-1,000 (who received blocks A and C). The correlations of items in block A with those in block B were estimated using examinees 1-333 only because no other subjects received both of these blocks.

One way to compare these two correlation matrices is in terms of their residual matrix, computed by subtracting the BIB matrix from the complete data matrix. Table 10.1(8) gives the lower quartile, median, and upper quartile of the distributions of several different types of residuals. The first two lines apply to the 30(29)/2 = 435 distinct off-diagonal elements of the residual matrix. Descriptive statistics are given for residuals (r_i) and for absolute residuals ($|r_i|$). The residuals were centered around zero; fifty percent of them were between -.02 and +.02. The median absolute residual was .02; fifty percent of the absolute residuals were between .01 and .04. The next two lines of Table 10.1(8) give the analogous information for residuals corresponding to within-block correlations (i.e., the diagonal blocks of Table 10.1(7)); the last two lines pertain to the residuals correponding to across-block correlations (i.e., the off-diagonal blocks of Table 10.1(7)). Because the within-block correlations were based on twice as many examinees as the across-block correlations, within-block residuals were smaller in absolute value.

Table 10.1(9) gives a partial comparison of the eigenstructures of the two correlation matrices. The lefthand side of the table shows the first ten eigenvalues of the two matrices; the righthand side gives ten elements of the first two normalized eigenvectors. Clearly, these eigenvalues and eigenvectors were very similar for the two matrices. Although subsequent eigenvectors were more discrepant, application of conventional

Table 10.1(7)

Subjects Available to Estimate Within- and Across-Block Correlations for 30-item BIB Simulation*

	Block A (Items 1-10)	Block B (Items 11-20)	Block C (Items 21-30)
A	667 Ss (1-333,667-1000)		
B	333 Ss (1-333)	666 Ss (1-666)	
C	334 Ss (667-100)	333 Ss (334-666)	667 Ss (334-1000)

*The table gives the number of subjects (Ss) available to estimate the correlations in each block of the matrix. The sequence numbers of the subjects are given in parentheses.

Table 10.1(8)

Distribution of Residual Correlations
for 30-Item BIB and Complete Data Simulations*

	Lower Quartile	Median	Upper Quartile
Full residual matrix (435 elements)			
Residuals	-.0236	-.0002	.0206
Absolute residuals	.0115	.0230	.0425
Within-block residual correlations (135 elements)			
Residuals	-.0165	-.0008	.0135
Absolute residuals	.0076	.0142	.0219
Across-block residual correlations (300 elements)			
Residuals	-.0299	-.0001	.0286
Absolute residuals	.0159	.0292	.0518

*Elements of the BIB matrix were subtracted from elements of the complete data matrix. Descriptive statistics were computed for:

(1) the 30(29)/2 = 435 distinct off-diagonal elements of the residual matrix;
(2) the 3[10(9)/2] = 135 distinct within-block residual correlations; and
(3) the $3(10^2) = 300$ across-block residual correlations.

Table 10.1(9)

Partial Comparison of Eigenstructure of BIB and Complete Data Correlation Matrices for 30-Item Simulation

		Ten Elements of Normalized Eigenvectors			
First Ten Eigenvalues		First Eigenvector		Second Eigenvector	
Complete	Incomplete	Complete	Incomplete	Complete	Incomplete
7.55	7.48	.20	.19	-.34	-.28
1.82	1.87	.22	.21	-.15	-.13
1.06	1.29	.17	.15	.11	.14
1.03	1.21	.21	.19	.07	.07
1.01	1.10	.10	.10	.19	.21
0.98	1.08	.16	.16	-.14	-.11
0.90	1.03	.14	.14	.25	.27
0.89	0.98	.19	.18	-.11	-.12
0.84	0.90	.18	.17	.24	.22
0.81	0.89	.21	.21	-.09	-.06

factor-analytic methodology to the BIB matrix would probably lead to conclusions that did not differ substantially from those obtained using the complete data correlation matrix. It is important to note however, that theoretical work is needed to fully understand the statistical properties of BIB matrices.

An important property of the full-information factor analysis and the Mantel-Haenszel approach is that they do not require the computation of the inter-item correlation matrix. That is, estimation of the parameters of interest in these models (factor loadings and item thresholds in full-information factor analysis, conditional odds ratios in the Mantel-Haenszel method) does not require an estimate of the population correlation matrix. The full-information factor analysis operates on the set of distinct vectors of item responses; the Mantel-Haenszel approach involves consideration of the pairwise relations between items. In neither case does the model theory dictate that the item response matrix be complete. This is a distinct advantage for NAEP applications. Essentially, the effect of the BIB missing data pattern on these analyses is that some parameters are estimated with greater precision than others. This uneven precision is unlikely to have a major effect on conclusions about dimensionality.

10.1.5 Conclusions

Overall, the four dimensionality analyses of the NAEP reading items indicate that it is not unreasonable to treat the data as unidimensional. As a preliminary approach, principal component analyses of phi and tetrachoric correlation matrices were computed for each of the three grade/ages and for the 25 across-grade/age items. The first roots obtained from these analyses were sizable, ranging from 17 to 25 percent of the trace for the phi matrices and 30 to 40 percent for the tetrachoric matrices. (For simulated unidimensional data, the first root of the phi matrix typically constituted 25 to 30 percent of the trace.)

As an experimental method, a factor-analytic approach based on Guttman's image theory was also applied. Principal component analysis of the image correlation matrices yielded larger first roots than PCA of the corresponding phi matrices, but larger second roots as well. However, both theoretical and empirical examinations of this method indicate that the image approach does not avoid the artifacts associated with the application of linear factor-analytic methods to dichotomous data.

Application of full-information factor analysis, a method developed by Bock and his associates, to a subset of the Grade 8/Age 13 data led to a satisfactory fit with a one-factor model. The first factor accounted for 39 percent of the total variance. Reading comprehension items involving fictional stories had the highest loadings on this factor; life skills items had the lowest.

Finally, the Mantel-Haenszel approach developed by Rosenbaum led to a retention of the hypothesis that the data can be represented by a unidimensional latent variable model with conditional independence. In addition to analyses within each grade/age, tests were conducted to determine whether data for each pair of grade/ages were consistent with a difference in distribution of a unidimensional latent variable. Again, the hypothesis of unidimensionality was retained.

Although categorization of the NAEP reading items is useful for test development and reading research, the dimensionality analyses reported here do not provide strong empirical evidence for the existence of multiple dimensions. Especially when considered in light of the robustness research discussed in Section 10.1.1.1, the results do not contraindicate the application of unidimensional item response theory models to the reading data.

Appendix 1

Items Used in Dimensionality Analyses

This appendix lists, for each grade/age, the items used in the NAEP dimensionality analyses. Items are listed by NAEP ID and by booklet location. (Note that the NAEP ID uniquely identifies an item. However, the booklet location for an item may differ across grade/ages.) The dimensionality analyses were given the codes A-E in this appendix. The following key explains these codes and indicates which section of the report contains an explanation of the analyses.

A. Component analysis and image analysis - within grade/age (see Sections 10.1.3.2 and 10.1.3.3)

B. Component analysis and image analysis - across grade/age (see Sections 10.1.3.2 and 10.1.3.3)

C. Full-information factor analysis (see Section 10.1.3.4)

D. Rosenbaum method - within grade/age (see Section 10.1.3.5)

E. Rosenbaum method - across grade/age (see Section 10.1.3.5.1)

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 4/Age 9

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
1.	N001101	H-005	X			X	
2.	N001501	H-010	X	X		X	X
3.	N001502	H-011	X	X		X	X
4.	N001503	H-012	X	X		X	X
5.	N001504	H-013	X	X		X	X
6.	N001506	H-015	X	X		X	X
7.	N001601	J-012	X				
8.	N001602	J-013	X				
9.	N001603	J-014	X				
10.	N001604	J-015	X				
11.	N001802	J-020	X				
12.	N002001	K-009	X	X		X	X
13.	N002002	K-010	X	X		X	X
14.	N002003	K-011	X	X		X	X
15.	N002101	K-018	X	X		X	X
16.	N002102	K-019	X	X		X	X
17.	N002401	L-022	X				
18.	N002702	L-020	X				
19.	N002801	L-024	X				
20.	N002802	L-025	X				
21.	N002803	L-026	X				
22.	N003001	M-010	X	X		X	X
23.	N003002	M-011	X	X		X	X
24.	N003003	M-012	X	X		X	X
25.	N003101	M-014	X	X		X	X
26.	N003102	M-015	X	X		X	X
27.	N003103	M-016	X	X		X	X
28.	N003701	N-023	X	X		X	X
29.	N003702	N-024	X	X		X	X
30.	N003703	N-025	X	X		X	X
31.	N003801	O-012	X	X		X	X
32.	N003802	O-013	X	X		X	X
33.	N003803	O-014	X	X		X	X
34.	N004101	O-017	X			X	
35.	N004201	O-018	X	X		X	X
36.	N004202	O-019	X	X		X	X
37.	N004401	P-007	X				
38.	N004402	P-008	X				
39.	N004403	P-009	X				
40.	N004701	Q-010	X				
41.	N004702	Q-011	X				
42.	N004703	Q-012	X				
43.	N004801	Q-013	X				
44.	N004901	Q-014	X	X			

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 4/Age 9

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
45.	N005101	Q-015	X				
46.	N008601	H-006	X			X	
47.	N008602	H-007	X			X	
48.	N008603	H-008	X			X	
49.	N008701	H-009	X			X	
50.	N008801	J-018	X				
51.	N008901	J-021	X				
52.	N008902	J-022	X				
53.	N008904	J-024	X				
54.	N009001	K-012	X			X	
55.	N009002	K-013	X			X	
56.	N009003	K-014	X			X	
57.	N009004	K-015	X			X	
58.	N009101	K-016	X			X	
59.	N009201	K-017	X			X	
60.	N009401	L-023	X				
61.	N009601	L-021	X				
62.	N009701	M-005	X			X	
63.	N009702	M-006	X			X	
64.	N009703	M-007	X			X	
65.	N009704	M-008	X			X	
66.	N009705	M-009	X			X	
67.	N009801	N-012	X			X	
68.	N009901	N-013	X			X	
69.	N010002	N-018	X			X	
70.	N010003	N-019	X			X	
71.	N010102	N-021	X			X	
72.	N010103	N-022	X			X	
73.	N010201	O-016	X			X	
74.	N010301	O-015	X			X	
75.	N010401	O-020	X			X	
76.	N010402	O-021	X			X	
77.	N010403	O-022	X			X	
78.	N010501	P-010	X				
79.	N010502	P-011	X				
80.	N010503	P-012	X				
81.	N010504	P-013	X				
82.	N010601	P-014	X				
83.	N010602	P-015	X				
84.	N010603	P-016	X				
85.	N010604	P-017	X				
86.	N010605	P-018	X				
87.	N010701	P-019	X				
88.	N010801	Q-016	X				
89.	N010902	Q-018	X				

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 4/Age 9

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
90.	N010903	Q-019	X				
91.	N010904	Q-020	X				
92.	N011001	R-005	X				
93.	N011002	R-006	X				
94.	N011003	R-007	X				
95.	N011004	R-008	X				
96.	N011101	R-009	X				
97.	N011201	R-010	X				
98.	N011301	R-011	X				
99.	N011302	R-012	X				
100.	N011401	R-013	X				
101.	N011402	R-014	X				
102.	N011403	R-015	X				
103.	N011404	R-016	X				
104.	N014001	M-013	X			X	
105.	N014101	Q-021	X				
106.	N014301	N-014	X			X	
107.	N014302	N-015	X			X	
108.	N014303	N-016	X			X	

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 8/Age 13

	NAEP ID	BOOKLET LOCATION	ANALYSES A	B	C	D	E
1.	N001101	H-006	X			X	
2.	N001201	H-007	X		X	X	
3.	N001202	H-008	X		X	X	
4.	N001301	H-009	X		X	X	
5.	N001302	H-010	X		X	X	
6.	N001303	H-011	X		X	X	
7.	N001401	H-012	X			X	
8.	N001501	H-013	X	X	X	X	X
9.	N001502	H-014	X	X	X	X	X
10.	N001503	H-015	X	X	X	X	X
11.	N001504	H-016	X	X	X	X	X
12.	N001506	H-018	X	X	X	X	X
13.	N001601	J-011	X		X		
14.	N001602	J-012	X		X		
15.	N001603	J-013	X		X		
16.	N001604	J-014	X		X		
17.	N001701	J-017	X		X		
18.	N001702	J-018	X		X		
19.	N001703	J-019	X		X		
20.	N001802	J-021	X		X		
21.	N001901	J-022	X		X		
22.	N001903	J-024	X		X		
23.	N002001	K-009	X	X		X	X
24.	N002002	K-010	X	X		X	X
25.	N002003	K-011	X	X		X	X
26.	N002101	K-012	X	X		X	X
27.	N002102	K-013	X	X		X	X
28.	N002201	K-014	X			X	
29.	N002202	K-015	X			X	
30.	N002203	K-016	X			X	
31.	N002902	M-006	X			X	
32.	N002903	M-007	X			X	
33.	N002904	M-008	X			X	
34.	N002905	M-009	X			X	
35.	N002906	M-010	X			X	
36.	N003001	M-011	X	X		X	X
37.	N003002	M-012	X	X		X	X
38.	N003003	M-013	X	X		X	X
39.	N003101	M-014	X	X		X	X
40.	N003102	M-015	X	X		X	X
41.	N003103	M-016	X	X		X	X
42.	N003201	N-012	X		X	X	
43.	N003202	N-013	X		X	X	
44.	N003203	N-014	X		X	X	
45.	N003204	N-015	X		X	X	

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
46.	N003301	N-016	X		X	X	
47.	N003401	N-017	X		X	X	
48.	N003501	N-018	X		X	X	
49.	N003601	N-019	X		X	X	
50.	N003602	N-020	X		X	X	
51.	N003701	N-021	X	X	X	X	X
52.	N003702	N-022	X	X	X	X	X
53.	N003703	N-023	X	X	X	X	X
54.	N003801	O-012	X	X	X	X	X
55.	N003802	O-013	X	X	X	X	X
56.	N003803	O-014	X	X	X	X	X
57.	N003901	O-016	X		X	X	
58.	N004002	O-015	X		X	X	
59.	N004101	O-017	X		X	X	
60.	N004201	O-018	X	X	X	X	X
61.	N004202	O-019	X	X	X	X	X
62.	N004301	O-020	X		X	X	
63.	N004302	O-021	X		X	X	
64.	N004401	P-007	X				
65.	N004402	P-008	X				
66.	N004403	P-009	X				
67.	N004501	P-010	X				
68.	N004502	P-011	X				
69.	N004601	P-012	X				
70.	N004602	P-013	X				
71.	N004603	P-014	X				
72.	N004604	P-015	X				
73.	N004701	Q-007	X				
74.	N004702	Q-008	X				
75.	N004703	Q-009	X				
76.	N004801	Q-010	X				
77.	N004901	Q-011	X	X			
78.	N005001	Q-013	X				
79.	N005002	Q-014	X				
80.	N005003	Q-015	X				
81.	N005101	Q-012	X				
82.	N005201	Q-016	X				
83.	N005202	Q-017	X				
84.	N005203	Q-018	X				
85.	N005301	Q-019	X				
86.	N005302	Q-020	X				
87.	N005303	Q-021	X				
88.	N005304	Q-022	X				
89.	N005305	Q-023	X				
90.	N005403	R-007	X				
91.	N005404	R-008	X				

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 8/Age 13

	NAEP ID	BOOKLET LOCATION	ANALYSES A	B	C	D	E
92.	N005405	R-009	X				
93.	N005406	R-010	X				
94.	N005407	R-011	X				
95.	N005503	R-014	X				
96.	N005504	R-015	X				
97.	N005505	R-016	X				
98.	N005601	R-017	X				
99.	N005602	R-018	X				
100.	N005603	R-019	X				

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 11/Age 17

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
1.	N001301	H-010	X			X	
2.	N001302	H-011	X			X	
3.	N001303	H-012	X			X	
4.	N001401	H-013	X			X	
5.	N001501	H-014	X	X		X	X
6.	N001502	H-015	X	X		X	X
7.	N001503	H-016	X	X		X	X
8.	N001504	H-017	X	X		X	X
9.	N001506	H-019	X	X		X	X
10.	N001701	J-012	X				
11.	N001702	J-013	X				
12.	N001703	J-014	X				
13.	N001901	J-015	X				
14.	N001903	J-017	X				
15.	N002001	K-009	X	X		X	X
16.	N002002	K-010	X	X		X	X
17.	N002003	K-011	X	X		X	X
18.	N002101	K-012	X	X		X	X
19.	N002102	K-013	X	X		X	X
20.	N002201	K-014	X			X	
21.	N002202	K-015	X			X	
22.	N002203	K-016	X			X	
23.	N002501	L-027	X				
24.	N002701	L-028	X				
25.	N002702	L-029	X				
26.	N002801	L-030	X				
27.	N002802	L-031	X				
28.	N002803	L-032	X				
29.	N002902	M-006	X			X	
30.	N002903	M-007	X			X	
31.	N002904	M-008	X			X	
32.	N002905	M-009	X			X	
33.	N002906	M-010	X			X	
34.	N003001	M-011	X	X		X	X
35.	N003002	M-012	X	X		X	X
36.	N003003	M-013	X	X		X	X
37.	N003101	M-014	X	X		X	X
38.	N003102	M-015	X	X		X	X
39.	N003103	M-016	X	X		X	X
40.	N003201	N-021	X			X	
41.	N003202	N-022	X			X	
42.	N003203	N-023	X			X	
43.	N003204	N-024	X			X	
44.	N003301	N-025	X			X	
45.	N003501	N-027	X			X	
46.	N003601	N-028	X			X	
47.	N003602	N-029	X			X	

ITEMS USED IN DIMENSIONALITY ANALYSIS - Grade 11/Age 17

			ANALYSES				
	NAEP ID	BOOKLET LOCATION	A	B	C	D	E
48.	N003701	N-030	X	X		X	X
49.	N003702	N-031	X	X		X	X
50.	N003703	N-032	X	X		X	X
51.	N003801	O-012	X	X		X	X
52.	N003802	O-013	X	X		X	X
53.	N003803	O-014	X	X		X	X
54.	N004201	O-021	X	X		X	X
55.	N004202	O-022	X	X		X	X
56.	N004301	O-023	X			X	
57.	N004302	O-024	X			X	
58.	N004501	P-020	X				
59.	N004502	P-021	X				
60.	N004601	P-022	X				
61.	N004602	P-023	X				
62.	N004603	P-024	X				
63.	N004604	P-025	X				
64.	N004901	Q-010	X	X			
65.	N005001	Q-007	X				
66.	N005002	Q-008	X				
67.	N005003	Q-009	X				
68.	N005201	Q-011	X				
69.	N005202	Q-012	X				
70.	N005203	Q-013	X				
71.	N005503	R-014	X				
72.	N005504	R-015	X				
73.	N005505	R-016	X				
74.	N015101	R-017	X				
75.	N015102	R-018	X				
76.	N015103	R-019	X				
77.	N015104	R-020	X				
78.	N015201	N-026	X			X	
79.	N015502	P-016	X				
80.	N015503	P-017	X				
81.	N015504	P-018	X				
82.	N015505	P-019	X				
83.	N015901	Q-014	X				
84.	N015902	Q-015	X				
85.	N015903	Q-016	X				
86.	N015904	Q-017	X				
87.	N016001	O-015	X			X	
88.	N016002	O-016	X			X	
89.	N016003	O-017	X			X	
90.	N016004	O-018	X			X	
91.	N016005	O-019	X			X	
92.	N016006	O-020	X			X	
93.	N017001	H-007	X			X	
94.	N017002	H-008	X			X	
95.	N017003	H-009	X			X	

Appendix 2

A Procedure for Obtaining a Gramian Matrix that Approximates a BIB Correlation Matrix for NAEP Items

(1) Start with the weighted (i.e., incorporating sampling weights) BIB covariance matrix.

(2) Substitute zeroes for the negative eigenvalues. (The negative eigenvalues constituted 4, 2, and 2 percent of the trace of the missing data covariance matrix for grade/ages 4/9, 8/13, and 11/17, respectively. There were no negative eigenvalues for the across-grade/age matrix.)

(3) Now obtain the "reconstructed" covariance matrix, $\underset{\sim}{C}^*$, using the following equation:

$$\underset{\sim}{C}^* = \underset{\sim}{Q}\,\underset{\sim}{D}^*\,\underset{\sim}{Q}',$$

where $\underset{\sim}{Q}$ is the matrix of normalized eigenvectors of the original covariance matrix and $\underset{\sim}{D}^*$ is a diagonal matrix of eigenvalues, with zeroes substituted for the negative eigenvalues. $\underset{\sim}{C}^{*-} = \underset{\sim}{Q}\,\underset{\sim}{D}^{*-1}\,\underset{\sim}{Q}'$ is the pseudo-inverse of $\underset{\sim}{C}^*$, where the elements of $\underset{\sim}{D}^{*-1}$ are the reciprocals of the corresponding elements of $\underset{\sim}{D}^*$ for positive elements of $\underset{\sim}{D}^*$ and zeroes for zero elements of $\underset{\sim}{D}^*$.

(4) It is now possible to obtain a reconstructed correlation matrix, $\underset{\sim}{R}^*$, corresponding to $\underset{\sim}{C}^*$, using ordinary methods. The pseudo-inverse of $\underset{\sim}{R}^*$ can be obtained as follows:

$$\underset{\sim}{R}^{*-} = \underset{\sim}{S}\,\underset{\sim}{C}^{*-}\underset{\sim}{S},$$

where $\underset{\sim}{S}$ is a diagonal matrix of the square roots of the diagonal elements of $\underset{\sim}{C}^*$.

It is desirable to begin with the covariance matrix in Step 1 because operating on the correlation matrix, $\underset{\sim}{R}$, directly will produce a reconstructed $\underset{\sim}{R}$ that does not have ones on the diagonal.

The medians of the residuals obtained by subtracting elements of $\underset{\sim}{R}^*$ from elements of the original $\underset{\sim}{R}$ were .007, .002, and .003 for grade/ages 4/9, 8/13, and 11/17, respectively. In addition, the eigenstructures for $\underset{\sim}{R}^*$ matrices were very similar to those for the original $\underset{\sim}{R}$'s. The method is inexpensive and is not difficult to program. An alternative method of B. Wingersky (1984) produced smaller residuals, but was prohibitively expensive to execute.

Chapter 10.2

JOINT ESTIMATION PROCEDURES

Marilyn Wingersky
Bruce A. Kaplan
Albert E. Beaton

Educational Testing Service

In its proposal for the NAEP grant (1982), ETS outlined how it would use the joint maximum likelihood procedures incorporated in the LOGIST program (see Wingersky, Barton, & Lord, 1982; M. S. Wingersky, 1983; M. S. Wingersky, 1984) to estimate reading item parameters and individual proficiencies. This method requires that a substantial number (25 or more) of exercises be administered to each student whose proficiency is to be estimated. Within the time available between receiving the grant and beginning field operation, the reading exercises, which were prepared by the previous grantee, could not be fitted into the block structure of the new design in such a way as to reach the number of exercises needed. The lack of sufficient exercises per student resulted in an undue number of students who had perfect scores or who scored below chance level and thus could not be assigned finite maximum likelihood estimates of their proficiencies. Because losing these students would bias population estimates made from the remaining data, winsorized estimates of the population parameters were computed. However, we then discovered a new technology that would provide better estimates, and thus this new method of parameter estimation was used.

The purpose of this section is to show the steps that we took in fulfilling the ETS commitment to use joint maximum likelihood procedures, the winsorization process, and the resultant effect on the distributions of reading proficiency. Chapter 10.3 will describe the new method of estimation that was actually used in producing the results that were presented in NAEP reports.

10.2.1 Method

The joint maximum likelihood estimation procedures incorporated in the LOGIST program are most appropriately applied to data sets in which each student responds to 25 or more exercises. Because the data collected in the Year 15 reading assessment did not meet the recommended minimum of 25 exercises per student, an alternative two-step estimation procedure was devised. In the first step of the alternative procedure, the LOGIST computer program was used to fit the three-parameter logistic IRT model to a sample of the available data. The sample was selected to maximize the precision of the estimated item parameters while minimizing convergence problems. In the second step of the procedure, the MLE-ABIL program was used to obtain maximum likelihood ability estimates (MLEs) for all students who were presented at least seventeen items. These steps are described in the following paragraphs. Additional details can be found in M. S. Wingersky (1986).

Both the LOGIST program and the MLE-ABIL program require an input data matrix consisting of observed item responses which have been coded as right, wrong, omitted or "not reached." The difference between an omitted response and a "not reached" response is described in Section 10.1.3.1.2. In brief, unanswered items which occur prior to the last valid response within a block are coded as omits. Unanswered items which occur subsequent to the last valid response in a block are coded as "not reached." In the Year 15 assessment, items marked "I don't know" were also coded as omits.

Both the LOGIST program and the MLE-ABIL program treat "not reached" items as if they had never been administered. The rationale for this treatment rests on a fundamental property of IRT models which states that an examinee's ability is invariant with respect to the items which are used to measure that ability. In the context of the NAEP reading assessment, this means that except for sampling fluctuations, an individual examinee's estimated reading proficiency value will be the same regardless of the particular subset of items to which the examinee has chosen to respond. Of course, reasonable numbers of responses are required to obtain precise parameter estimates. In the Year 15 calibration, a cutoff value of seventeen items per examinee was established. In the first step of the calibration, this cutoff was applied to the number of items reached. Omitted responses were included in the count of items reached. In the second step of the calibration, the same cutoff value of seventeen items was applied to the number of items presented. (The MLE-ABIL program can accept a slightly less stringent data input requirement because it is only estimating abilities; item parameters are fixed rather than estimated.)

Many applications of IRT allow for the fact that some examinees will respond correctly to an items by guessing. It is typically assumed that if an examinee elects to guess, the probability that he or she will guess correctly can be approximated by the reciprocal of the number of valid response alternatives. Both LOGIST and MLE-ABIL incorporate this assumption by maximizing a likelihood function which has been modified to allow partial credit for omitted responses. In effect, omitted responses are treated as fractionally correct, at a proportion equal to the reciprocal of the number of valid response alternatives. This modification is described in detail by Lord (1974).

10.2.2 Item Parameter Calibration

Initially the calibrations were done separately by grade/age to see how similar the parameter estimates were for the items that were common across the grade/ages. If the estimates were similar enough, all of the grade/ages could be calibrated together giving better parameter estimates for the common items and a better linking between the ages than if the ages were calibrated separately and linked with some standard linking procedure.

The first grade/age to be analyzed was Grade 8/Age 13, the middle grade/age in proficiency. Included in the calibration run were examinees who took two or more of blocks H, J, K, M, N, O, P, Q, R, and U and reached at least seventeen items regardless of how many items they omitted. Excluded from the calibration run were examinees who took blocks L, V, W, and Y, which had fewer than seven reading items. Although it would have been possible to calibrate these items when given with other blocks, final proficiencies estimated for examinees who took only these blocks would be poorly estimated because of the small number of items. Block X was also not included, even though it had eight reading items, because six of the items were puns, the only puns in the entire item collection. Examinees who had zero or perfect scores were excluded. Of the 10,255 examinees, 490 were removed because they reached fewer than seventeen items. There were 113 items and 9,765 examinees in the calibration run.

The same criteria used to determine which Grade 8/Age 13 examinees to include in the calibration were used for Grade 4/Age 9. Blocks included were H, J, K, L, M, N, O, P, Q, R, U, and V. Of the 13,297 examinees, 1,786 who reached fewer than seventeen items were removed. There were 127 items and 12,141 examinees in the calibration run. The Grade 4/Age 9 item parameter estimates were then transformed so that they would be on the same proficiency scale as the Grade 8/Age 13 item parameter estimates. The transformation program used, TBLT, computes the linear transformation that minimizes the squared difference between the two test characteristic curves computed for the common items (Stocking & Lord, 1983).

For the Grade 11/Age 17 calibration, it was necessary to include blocks with as few as five items. Otherwise, there would be too few examinees per item to calibrate any items. Thus blocks L and Y with only six items and block J with only five items were included. The blocks used were H, J, K, L, M, N, O ,P, Q, R, U and Y. Examinees who reached fewer than seventeen items were excluded. Booklets 2, 19, 27, 30, and 57 were excluded because they contained fewer than seventeen reading items, even though they contained two of the above reading blocks. Of the 12,011 examinees, 1,314 were removed because they reached fewer than seventeen items. There were 113 items and 10,697 examinees in the calibration run. The item parameters were then transformed using TBLT to the proficiency scale of Grade 8/Age 13.

The parameter estimates for the items that were common across the different grade/ages were consistent enough to warrant calibrating all of the grade/ages together. The same examinees and the same items for each grade/age were used as were used in the single calibrations. Although blocks L and Y were calibrated for Grade 11/Age 17 in the single run, they were not calibrated for Grade 8/Age 13. Consequently, the responses of Grade 8/Age 13 examinees to the items in blocks L and Y were coded as "not reached" for this run. Although the dataset of item responses contained 252 items, only 229 items were calibrated. Items in blocks W and X were not used and were coded "Not reached". Items in block V for Grade 8/Age 13 and Grade 11/Age 17 were not used and were coded "Not reached". Of the initial 36,193 examinees, 3,590 were removed because they reached fewer than seventeen items. There were 229 items and 32,603 examinees in the calibration run. Again, only examinees who reached seventeen items were included. Examinees were included regardless of the number of items coded as Omits.

Item proficiency regression plots, where the observed proportion correct were plotted separately for the three groups, were examined to see if the different grade/age groups were responding differently on the common items. Although there were several items for which individual grade/age groups responded differently, it was decided to use the results with all grade/ages calibrated together.

10.2.3 Maximum Likelihood Estimates of Proficiency

A student's maximum likelihood estimate (MLE) of proficiency on a given scale indicates the value that is most likely to have produced the responses that he or she actually made given the estimated item parameters for the items taken. A maximum likelihood estimate of θ cannot be computed for examinees with zero scores, perfect scores or a small number of other response patterns where the MLE of θ attempts to go to minus infinity.

Proficiency estimates were computed for all examinees for all grade/ages who took booklets that had seventeen or more reading items calibrated for that particular grade/age. Item N001801, which had a flat item response function, was not used. The proficiency range was bounded by -7 and 5. The "direct" method refined by several Newton iterations was used. The direct method consists of computing the likelihood function for equally-spaced proficiencies between -7 and 5 and selecting the proficiency corresponding to the maximum of these values of the likelihood function. This proficiency is then used as a starting value for Newton's method.

The standard error of estimation of the maximum likelihood estimate of proficiency was computed for each MLE. The standard error of estimation indicates the precision of measurement of the maximum likelihood estimate.

A score on the xi scale was also computed for each examinee. This is a nonlinear transformation of the θ scale to a number-right true score scale. The xi scale refers to the number of correct responses that might be expected if the entire pool of 228 reading exercises included in the IRT scaling were administered as a single test. This scale runs from 48.7 ("chance" level) to 228 (perfect score).

The proficiency estimate was also converted to a reading proficiency (RP) score. This refers to the expected number of correct responses that the examinee would get on a hypothetical 500-item test. The IRT parameters for the items on this test have equal item discriminations, at the average level of the actual NAEP items on the scale; equal lower asymptotes of zero; and equally-spaced difficulty parameters ranging from -5 to +4.98 on the proficiency scale. The relationship between this scale and the θ scale is virtually linear for proficiencies from -4 to +4 and is approximated by the relationship

$$\text{Proficiency} = 50\,\theta + 250.5$$

The standard errors of both xi and reading proficiency scores were computed. MLE proficiency estimates on the xi scale and the reading proficiency scale were computed for the same examinees for which proficiency estimates on the θ scale were computed.

The following bounds were placed on the values for the various scales. For examinees whose proficiencies could be estimated, the scale was limited to a range of -7 to 5. The standard error of θ was limited to a maximum of 998. The xi scale was limited to 48.7 to 228. The standard error of xi was limited to a maximum of 228. The proficiency scale was limited to 1 and 449 (the score to which a θ of 4 converts on the proficiency scale). The maximum standard error on the proficiency scale was 500.

Table 10.2(1) indicates the arbitrary values flagging examinees for whom a maximum likelihood estimate could not be obtained.

Because of the short length of the tests and the wide range of abilities spanning Grades 4 to 11, there were many zero and perfect scores. These scores at the extremes distorted statistics computed for different subgroups but they could not be dropped without destroying the representativeness of the sample to the population. Consequently a procedure that is a type of Winsorizing (Huber, 1981) was devised, to bring these extreme values closer to the rest of the values in the distribution.

This was done by computing the "inner fences" (Tukey, 1977, p. 44) as boundaries to the distribution and setting all values outside of the boundaries at the appropriate boundary. The boundaries were computed by first computing the hingespread, H, which is the difference between the 25th percentile (Q_1) and the 75th percentile (Q_3), then computing the minimum and maximum boundary (or inner fences) as follows:

$$H = Q_3 - Q_1$$

$$\text{minimum} = Q_1 - 1.5\ H$$

$$\text{maximum} = Q_3 + 1.5\ H$$

All values below the minimum were set to the minimum; all values above the maximum were set to the maximum.

Table 10.2(2) shows the minimum and maximum scores and the number of values changed for each grade/age for the Year 15 BIB spiral data.

Table 10.2(1)

Values Assigned to Examinees Whose Maximum Likelihood Estimate Could Not Be Computed

	θ	Flag for standard error (θ)	xi	Flag for standard error (xi)	RP score	Flag for standard error (RP)
Zero score	-100.	999.	48.7	999.	1.	999.
MLE below lower limit	-100.	999.	48.7	999.	1.	999.
Perfect score	100.	999.	228.0	999.	449.	999.

Table 10.2(2)

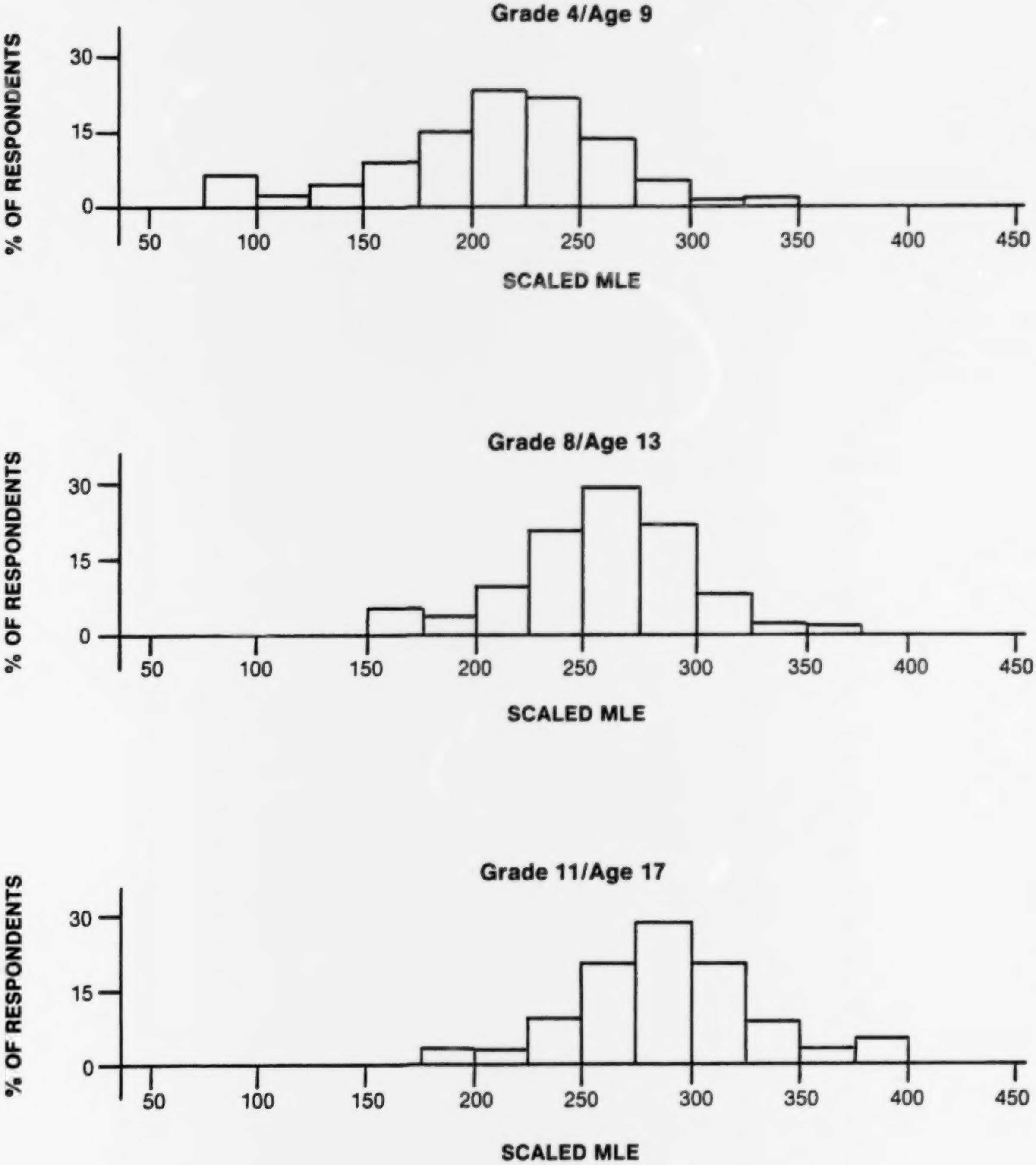
Minimum and Maximum Scores and Number of Values Changed by Grade/Age

Grade/ Age	Percent Moved to Minimum	Minimum	Q_1	Median	Q_3	Percent Moved to Maximum	Maximum
4/9	6.4	82	177	212.5	241	336	1.0%
8/13	4.5	161	233	257.5	281	353	1.2%
11/17	2.9	189	262	287.0	311	385	4.0%

The groupings of extreme scores, even after this modification, produce the distributions shown as Figure 10.2-1. The proportions of a grade/age population accounted for in extreme groups depends in part upon features unrelated to the true distributions, such as the numbers and difficulties of exercises administered to pupils. Doubts thus arise about using these distributions of estimates to approximate characteristics of the distribution of underlying proficiencies, where no such anomalies are anticipated. Methods intended to estimate the underlying distribution directly, bypassing the intermediate and problematic step of estimating scores for individual examinees, are described in Chapter 10.3.

Figure 10.2-1

Distributions of Adjusted Proficiency Scale Scores



Chapter 10.3

MARGINAL ESTIMATION PROCEDURES

Robert J. Mislevy
Kathleen M. Sheehan

Educational Testing Service

Item response theory (IRT) offers NAEP the advantages of efficiency in the estimation of population characteristics, common-scale measurement across forms and over time, and results that are interpretable in terms of expected behavior on specific tasks. The experiences described in the preceding chapter proved, however, that these advantages could not be attained with standard IRT measurement procedures. The NAEP data are simply too sparse at the level of the individual examinee to support precise individual point estimates--estimates which could be used in turn to estimate parameters for cognitive items, population characteristics, and relationships between performance and background variables.

However, it is exactly these latter population-level parameters, rather than parameters for specific examinees, that are of interest in NAEP. NAEP objectives can therefore be attained with methodologies that produce population parameters directly, without the intermediary computation of parameters for individuals. To this end, marginal estimation techniques for latent variables (e.g., Bock & Aitkin, 1981; Mislevy, 1985a) were extended to the setting of survey samples by means of Rubin's (1977, 1978) multiple imputation techniques for missing data. A technical description of the resulting procedure is given in Mislevy (1985b). The purposes of this chapter are to (1) review the procedures in general terms and (2) provide details of their implementation in the Year 15 NAEP reading analyses. The steps in those analyses, which will be discussed in turn after an overview of the procedures, are as follows:

Year 15 BIB Data

- Estimation of item parameters
- Estimation of conditional effects
- Generation of "plausible values"

Year 15 Pace Data

- Estimation of conditional effects
- Generation of "plausible values"
- Equating to BIB scale

(The estimation of item parameters under paced administration conditions was carried out in conjunction with the analysis of data from previous NAEP reading assessments, and will be described in Chapter 10.4, Estimation of Trends.)

10.3.1 The General Model

The object of inference in a sample survey is a (possibly vector-valued) function T of the values of survey variables in all N members of the population. This value is estimated by a function t of the values obtained from a sample of size n. The precision of t as an estimate of T is indicated by another function of the sampled values, namely the estimated variance var(t), which approximates the true variance of t, or VAR(t).

To enable discussion, we shall denote the (possibly vector-valued) proficiency of examinee i by θ_i, and denote by $\underset{\sim}{\theta}$ the values in the entire realized sample. Let y_i and $\underset{\sim}{y}$ denote similarly defined values of background and attitude variables for examinee i and for the entire realized sample respectively. If $\underset{\sim}{\Theta}$ and $\underset{\sim}{Y}$ represent correspondingly defined values in the population as a whole, then T is a function of $\underset{\sim}{\Theta}$ and $\underset{\sim}{Y}$, while t and var(t) are functions of $\underset{\sim}{\theta}$ and $\underset{\sim}{y}$.

The formulations above assume that $\underset{\sim}{\theta}$ is observed without error. This is not the case in NAEP under the assumed IRT model. Instead, observations are of the form x_{ij}, the response of examinee i to cognitive item j, for $j=1,\ldots,m$. These responses are assumed to be governed by the IRT model, under which the probability of a given response depends on the (unobserved) proficiency of the examinee and the (unknown) parameters β_j of the item through the IRT function $p(x_{ij} = 1|\theta_i,\beta_j)$.

A latent variable like θ in an IRT model can be thought of as a variable whose value is missing for all examinees. Under Rubin's (1977) approach to missing values in survey samples, a reasonable estimate of T is obtained by computing the expectation of t, given values of variables that were not missing, i.e.,

$$t^*(\underset{\sim}{x},\underset{\sim}{y}) = \underline{E}(t(\underset{\sim}{\theta},\underset{\sim}{y})|\underset{\sim}{x},\underset{\sim}{y})$$

$$= \int t(\underset{\sim}{\theta},\underset{\sim}{y})\ p(\underset{\sim}{\theta}|\underset{\sim}{x},\underset{\sim}{y})\ d\underset{\sim}{\theta} \quad . \qquad (1)$$

Equation (1) may be thought of as an average of $t(\underset{\sim}{\theta},\underset{\sim}{y})$ computed over all possible values of the unobserved variable $\underset{\sim}{\theta}$, with each weighted in proportion to its consonance with the observed values $\underset{\sim}{x}$ and $\underset{\sim}{y}$. Furthermore, the variance of t* can be approximated by

$$\text{var}(t^*) = \underline{E}(\text{var}(t(\underset{\sim}{\theta},\underset{\sim}{y}))|\underset{\sim}{x},\underset{\sim}{y}) + \text{Var}(t(\underset{\sim}{\theta},\underset{\sim}{y})|\underset{\sim}{x},\underset{\sim}{y}) \qquad (2)$$

(Hertzog & Rubin, 1983). This variance estimator is the sum of two components: the expected value of the variance of t, which indicates uncertainty due to sampling from the population, and the variance of $t(\underset{\sim}{\theta},\underset{\sim}{y})$ given $\underset{\sim}{x}$ and $\underset{\sim}{y}$, which indicates uncertainty due to not knowing the θ values of the examinees in the realized sample.

The evaluation of Equations (1) and (2) requires the conditional distribution of the latent variables $\underset{\sim}{\theta}$ given the observed variables $\underset{\sim}{x}$ and $\underset{\sim}{y}$, or $p(\underset{\sim}{\theta}|\underset{\sim}{x},\underset{\sim}{y})$. Standard rules of the calculus of probabilities allow this distribution to be expressed as a constant times the product of two terms, or

$$p(\underset{\sim}{\theta}|\underset{\sim}{x},\underset{\sim}{y}) \propto p(\underset{\sim}{x}|\underset{\sim}{\theta},\underset{\sim}{y})\, p(\underset{\sim}{\theta}|\underset{\sim}{y}) \quad . \qquad (3)$$

The first term in the right hand side of this expression is given by the item response model. By conditional independence,

$$p(\underset{\sim}{x}|\underset{\sim}{\theta},\underset{\sim}{y}) = \prod_i \prod_j p(x_{ij}|\theta_i,\beta_j)$$

where again β_j is the unknown and possibly vector-valued parameter for item j. If multiple scales pertaining to mutually-exclusive subsets of items are entertained, this term may be written as

$$p(\underset{\sim}{x}|\underset{\sim}{\theta},\underset{\sim}{y}) = \prod_i \prod_k \prod_j p(x_{ijk}|\theta_{ik},\beta_{ik}) \quad ,$$

where k indexes scales and β_{jk} is the parameter for item j within scale k.

Assuming independence over examinees, the second term in Equation (3) can be written as

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \prod_i p(\theta_i|y_i,\underset{\sim}{\alpha}) \quad ,$$

the product over examinees of conditional distributions of the values of their latent variables given their observed responses to background and attitude items. Here $\underset{\sim}{\alpha}$ represents the (unknown) parameters of these distributions. Suppose, for example, normal distributions are assumed for conditional distributions whose means are determined by y. In this case $\underset{\sim}{\alpha}$ might consist of a common conditional variance and regression parameters that yield the conditional means.

The unknown parameters $\underset{\sim}{\beta}$ for item parameters and $\underset{\sim}{\alpha}$ for conditional distributions of θ given background variables y can be estimated precisely from large samples of examinees even when individual examinees' parameters cannot. This may be accomplished by so-called "marginal" estimation procedures that, in statistical terms, treat examinee parameters as random rather than fixed effects. Both sets of parameters may be estimated simultaneously by the method of maximum likelihood, for example, by maximizing the following marginal likelihood function with respect to $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$:

$$L(\underset{\sim}{\alpha},\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y}) = \prod_{i} \int p(x_i|\theta,y_i,\underset{\sim}{\beta})\, p(\theta|y_i,\underset{\sim}{\alpha})\, d\theta \quad . \qquad (4)$$

An algorithm to accomplish this task is given in Mislevy (1986a). This algorithm was applied to the Year 15 reading data in two steps. First, the vector of items parameters $\underset{\sim}{\beta}$ was estimated with respect to an unrestricted θ distribution. Second, the conditional effects $\underset{\sim}{\alpha}$ were estimated with $\underset{\sim}{\beta}$ fixed at its maximizing value. The first step was accomplished using BILOG (Mislevy & Bock, 1982). The second step was accomplished using M-GROUP (Sheehan, 1985).

The parameter estimates $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\alpha}}$ were then used to approximate the conditional distribution of θ given x and y, for each examinee, as follows:

$$p(\theta|x_i,y_i) \approx p(\theta|x_i,y_i,\underset{\sim}{\alpha}=\hat{\underset{\sim}{\alpha}},\underset{\sim}{\beta}=\hat{\underset{\sim}{\beta}}) \qquad (5)$$

where $\hat{\underset{\sim}{\beta}}$ = estimated item parameters obtained from step 1, and

$\hat{\underset{\sim}{\alpha}}$ = estimated conditional effects obtained from step 2.

10.3.1.1 "Plausible Values"

Two considerations merit attention at this point. First, even when point estimates of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ are used to approximate $p(\theta|x,y)$ in the manner described above, neither closed-form expressions nor convenient analytic approximations are generally available; instead, numerical approximations must be employed. Second, it is not possible in a survey with as many background variables as the NAEP survey to model in detail the full conditional distribution $p(\theta|y)$; only selected background variables can be included, and even then, a simplified functional form must be used. These considerations are discussed, in turn, below.

The numerical approximations employed by M-GROUP can be characterized by (a) the representation of smooth functions such as $p(x_i|\theta,y_i)$, $p(\theta|y_i)$, and $p(\theta|x_i,y_i)$ as histograms over points that span the range of θ, and (b) Monte Carlo evaluation of required integrals via repeated samples from $p(\theta|x,y)$. Each histogram is defined over a predetermined grid of points $\underset{\sim}{A}$. $p(\theta|x_i,y_i)$ is then approximated, at each point A_q in A, as

$$P(\theta = A_q | x_i, y_i) = C_i \; p(x_i | A_q, \hat{\underset{\sim}{\beta}}) \; p(A_q | y_i, \hat{\underset{\sim}{\alpha}}) \quad , \qquad (6)$$

where C_i is a normalizing constant. A value $\tilde{\theta}_i$ may then be drawn at random from the histogram in two steps. First, a bar is selected at random from the histogram in accordance with the probabilities given by Equation (6). Second, a value is selected at random from that interval. Carrying out these steps for each examinee in the sample yields a pseudo-dataset, with each examinee represented by a "plausible value" of what his or her unobservable θ might be, given the observed values x_i and y_i.

This construction guarantees that $t(\tilde{\underset{\sim}{\theta}},y)$ and $var(t(\tilde{\underset{\sim}{\theta}},y))$ have expectations equal to $\underline{E}(t(\underset{\sim}{\theta},\underset{\sim}{y})|\underset{\sim}{x},\underset{\sim}{y})$ and $\tilde{\underline{E}}(var(t(\underset{\sim}{\theta},\underset{\sim}{y}))|\underset{\sim}{x},\underset{\sim}{y})$, the values targeted in Equations (1) and (2). Let $\tilde{\underset{\sim}{\theta}}_k$ represent the vector of plausible values comprising the k^{th} of K pseudo-datasets. Under the assumption that $p(x|\theta)$ and $p(\theta|y)$ have been correctly specified, a consistent estimate of T is given by

$$t^* = K^{-1} \sum_k t(\tilde{\underset{\sim}{\theta}}_k, \underset{\sim}{y}) \quad .$$

Its variance, as an estimate of T, may be approximated as

$$var(t^*) = K^{-1} \sum_k var(t(\tilde{\underset{\sim}{\theta}}_k, \underset{\sim}{y})) + (K-1)^{-1} \sum_k [t^* - t(\tilde{\underset{\sim}{\theta}}_k, \underset{\sim}{y})]^2 \quad .$$

This variance estimator is again the sum of two terms. The first, representing sampling variability, may take the form of jackknifing t with $\tilde{\underset{\sim}{\theta}}$ treated as an observed variable; alternatively, a simple random sampling variance, again evaluated on $\tilde{\theta}$, may be boosted by a design effect. The second term again reflects uncertainty due to the latency of θ .

10.3.1.2 Effects of Specification Errors on Plausible Values

It is implicit in Equation (3) that consistent estimation of statistics involving background variables requires that the joint density of those variables with the unobservable variable be specified and its parameters estimated. It is obviously not possible to compute a joint distribution for all of the hundreds of NAEP variables; the procedures were employed for only the key NAEP reporting variables for trends:

- Age
- Grade
- At, above, or below modal age for grade

- At, above, or below modal grade for age
- Sex
- Ethnicity (Hispanic, black, and other)
- Size and type of community (high metro, low metro, and other)
- Parental education (higher of mother or father: less than high school, high school, or post high school)
- Region of the country

We shall refer to these as the "conditioning variables." Moreover, only a main effects model, assuming normally distributed and homoscedastic residual terms, could be employed due to computational limitations. The distribution $p(\theta|y)$ is thus approximated by $p^*(\theta|y)$, where p^* incorporates only main effects of the above-mentioned conditioning variables.

This simplification can be considered a "primary" specification error, "primary" because it enters into the generation of plausible values. It is distinguished from a "secondary" specification error, which would refer to, say, omitting variables from a regression equation when analyzing a given set of plausible values. The consequences of primary specification error for subsequent analyses can be expressed as follows:

$$\text{Bias} = \underline{E} \left\{ \int t(\underset{\sim}{\theta},\underset{\sim}{Y}) \, [p^*(\underset{\sim}{\theta}|\underset{\sim}{X},\underset{\sim}{Y}) - p(\underset{\sim}{\theta}|\underset{\sim}{X},\underset{\sim}{Y})] \, d\underset{\sim}{\theta} \right\} , \qquad (7)$$

where expectation is taken over $\underset{\sim}{X}$ for fixed $\underset{\sim}{\theta}$. Of particular interest in NAEP are biases corresponding to nonconditioned variables. That is, $Y=(Y_1,Y_2)$, plausible values are generated under $p^*(\theta|Y) = p(\theta|Y_1)$, and a statistic t that involves θ and Y_2 is calculated under a secondary analysis.

Unfortunately, simple expressions for (7) are not readily available in full generality for all the statistics that could be computed from the NAEP database. Section 10.3.5 instead presents primary specification biases in a simplified case for which explicit expressions can be derived. Specifically, we shall assume a variant of the classical "true-score" model of test theory, under which the variable $X=\theta + E$ is observed (along with Y), with E independent and identically distributed normal over all respondents. Note that no such X can be calculated for θ in the NAEP setting of an IRT model with only a few responses per person. This simplified setting does, however, provide both intuition and approximate expressions for the more complex relationships between latent and observed variables that are embedded in IRT. Details are given in Mislevy (in progress).

10.3.2 Estimation of Item Parameters

The LOGIST computer program (Wingersky, Barton, & Lord, 1982) had originally been used to obtain estimated item parameters for 229 of the reading items which were administered in the Year 15 assessment. However, because the LOGIST results proved to be unsatisfactory, item parameters were re-estimated using the BILOG computer program (Mislevy & Bock, 1982). In both calibrations a three-parameter logistic IRT model was assumed.

Like the LOGIST program, BILOG requires an input data matrix consisting of observed item responses which have been coded as right, wrong, omitted or "not reached." The coding conventions developed for the LOGIST calibration were used, without modification, for the BILOG calibration. (These coding conventions are described in Section 10.2.1.) The BILOG calibration also mirrored the LOGIST calibration in its treatment of omitted and "not reached" responses. For the reasons presented in Section 10.2.1, responses coded as "not reached" were excluded and responses coded as "omitted" were treated as fractionally correct, at a proportion equal to the number of valid response alternatives.

The major difference between the LOGIST calibration and the BILOG calibration is that the joint estimation procedures employed by LOGIST require that a point estimate of proficiency be computed for each subject, whereas the marginal estimation procedures employed by BILOG rely on Bayes Theorem to obtain proficiency distributions for all subjects without computing individual proficiency point estimates for any subject.

In both programs, estimation proceeds in cycles, with provisional proficiency estimates (or distributions) used to obtain improved item parameter estimates in one cycle and provisional item parameter estimates used to obtain improved proficiency estimates (or distributions) in the next cycle.

A practical result of the differences noted between the estimation procedures employed by BILOG and LOGIST is that BILOG does not require that each examinee respond to a minimum number of items. Instead, BILOG's data input requirements are formulated in terms of the number of examinees responding to each item. In particular, it is recommended that each calibration be performed on a data set providing a minimum of 1,000 responses for each item.

Because IRT parameters are theoretically sample-free, and IRT calibration programs are generally expensive to run, many IRT models are calibrated from a sample of the available data. The calibration sample is typically selected to meet, but not exceed, the input data requirements of the particular calibration program being used. When the calibration sample is randomly selected from the available data, resulting parameter estimates are unbiased estimates of those that would have been obtained if all of the data had been used, as long as all of the data input requirements have been satisfied. The invariance property of IRT item parameters also provides the theoretical justification for not using sampling weights during the item parameter estimation phase of an IRT calibration.

Each of the 229 items which had been selected for use in the LOGIST calibration were considered for inclusion in the BILOG calibration. All but one were eventually included. Item 20 (ETS ID # N001201) was excluded because it had exhibited severe convergence problems. The calibration sample was selected from all examinees who took at least two reading blocks, except for excluded blocks. (Blocks W and X were excluded for Grade 4/Age 9. Blocks V, W, and X were excluded for Grade 8/Age 13 and Grade 11/Age 17.) This BILOG sampling frame differed from the LOGIST sampling frame in that examinees in Grade 8/Age 13 who took blocks L or Y were not excluded. The BILOG and LOGIST sampling frames are summarized in Table 10.3(1).

The final sample consisted of 10,286 examinees, or approximately one fourth of the available subjects in each grade/age. This sample provided approximately 1,000 examinees in each grade/age for each item. However, since all items were not administered to all grade/ages, the total number of examinees responding to each item ranged from 1,000 to 3,000 (approximately). As noted above, sampling weights were not employed in the item calibration.

Several modifications were made to the BILOG computer program to customize it for use with NAEP data. One modification provided an option for analyzing items with variable numbers of response alternatives. A second modification provided a capability for distinguishing among distinct subpopulations of examinees in the calibration sample. This capability was required to avoid the gratuitous assumption that examinees in different grade/ages were exchangeable members of a common population. A final modification provided for the creation of an output file containing item fit statistics for subpopulations of examinees.

Although the three-parameter model has been shown to be well suited for analyzing NAEP data, it does have some unfortunate characteristics. One of these is a tendency to produce multi-collinearity when the response data includes very difficult or very easy items. In cases of multi-collinearity, widely varying combinations of the (a,b,c) parameters can produce similar response curves through the region of θ where the calibration sample of examinees lies. Without constraints, unstable and unreasonable (a,b,c) triples can result. BILOG guards against these problems by supplying Bayesian priors for each type of item parameter, with fixed dispersions and with locations estimated from the data. Default priors are normal for b's, with a standard deviation of 2; log-normal for a's with a standard deviation of 1 for log a; and beta for c's, with the weight of 20 observed responses from low-ability examinees. These default priors proved to be unsatisfactory for the multiple-choice items in the Reading assessment, primarily because of the presence of a large number of very easy items. In particular, estimated c values tended to be higher than expected (when compared with the reciprocals of the numbers of

Table 10.3(1)

Blocks Selected for Scaling the Year 15 Reading Data

Block	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
H	X	X	X
J	X	X	X
K	X	X	X
L	X	0	X
M	X	X	X
N	X	X	X
O	X	X	X
P	X	X	X
Q	X	X	X
R	X	X	X
U	X	X	X
V	X	N	N
W	N	N	N
X	N	N	N
Y	N	0	X

X = Included in both the LOGIST and BILOG item calibrations.
0 = Included in the BILOG item calibration only.
N = Not included in either calibration.

response alternatives) and estimated a's were lower than expected (when compared with a values from free-response items). To force the program to produce "more reasonable" estimates, the prior distributions were modified in the following manner:

(1) The prior standard deviation of log a was changed from 1.0 to 0.5, and

(2) the precision of the beta prior on asymptotes was increased from the weight of 20 observations to the weight of 50 observations.

These changes resulted in item parameter estimates that were reasonable in appearance and fit the data well. The resulting item parameter estimates and corresponding standard errors are provided in Appendix B, Table B-8. Because a linear indeterminacy exists with respect to the values of θ, a, and b, in the three-parameter model, the parameter estimates have been arbitrarily scaled so that the distribution of proficiency in the calibration sample has a mean of zero and a standard deviation of one.

Item fit was evaluated by inspecting residuals from fitted item response curves. A typical plot is shown in Figure 10.3-1. The smooth line is the fitted three-parameter logistic item response curve; the three plot symbols represent the expected proportions of correct responses for examinees in each grade/age at various points along the reading proficiency scale. These expected proportions were calculated without assuming the three-parameter functional form. (These plots were produced by a special modification of BILOG. Each is based on pseudo-counts of attempts and corrects to an item produced by an additional E-step of its EM-algorithm. See Bock and Aitkin, 1981, for details.) The size of each symbol is proportional to the amount of information available in the calibration dataset in the region of the scale where the symbol is plotted.

Item bias was evaluated by inspecting residuals for examinee subpopulations defined by sex and ethnicity. (These plots were produced by a special modification of LOGIST. Each is based on counts of attempts and corrects to an item from groups of examinees with similar estimated abilities.) A typical plot is shown in Figure 10.3-2. In this figure, the plot symbols distinguish between subpopulations defined by sex. Plots such as those depicted in Figures 10.3-1 and 10.3-2 were examined for all items. Copies of these plots are available from ETS upon request.

10.3.3 Estimation of Conditional Effects

Conditional distributions of reading proficiency given background responses were estimated separately for examinees in each grade/age. The number of background variables which could be included in each within-grade/age model was limited by the availability of computing resources. The background variables selected included sex, imputed ethnicity, size and type of community (STOC), region, and parental

Figure 10.3-1

Diagnostic Fit Plot for Item 9 (ID=N001502)

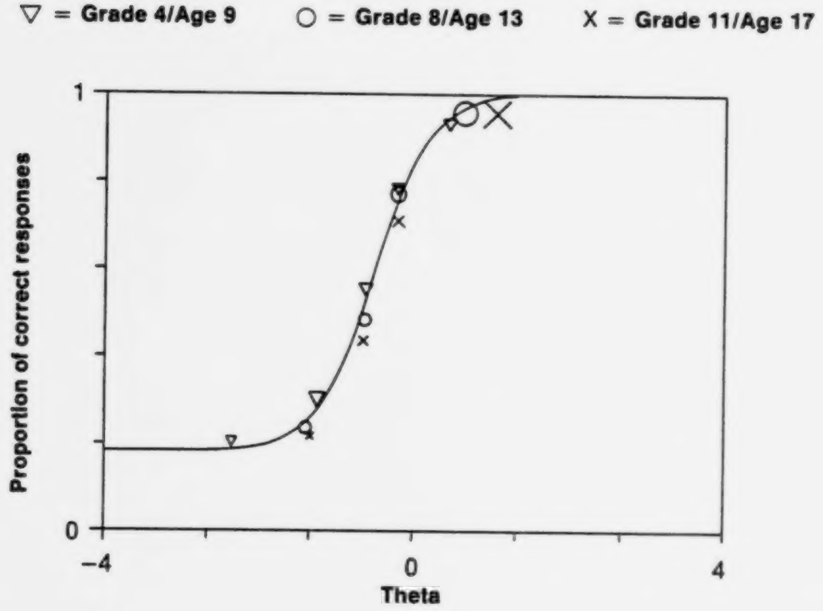
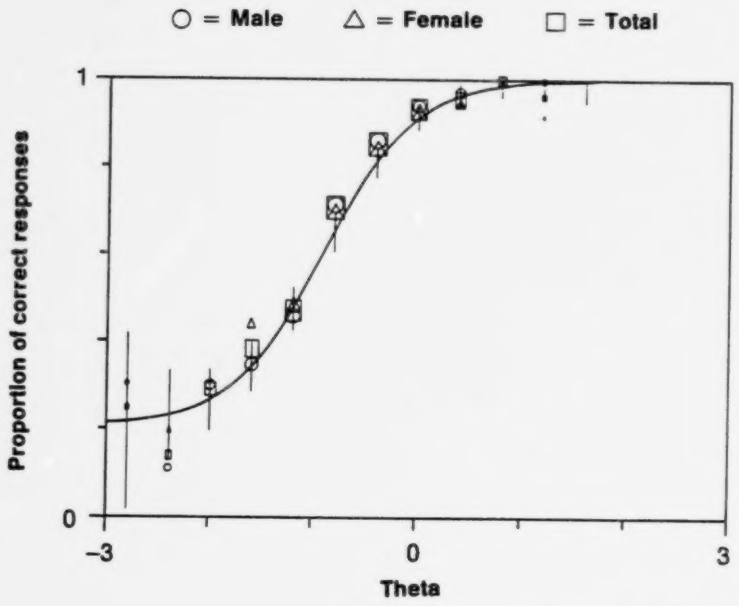


Figure 10.3-2

Bias Plot for Item 10 (ID=N001503)



education (these variables are defined in Section 12.1). Differences in reading proficiency resulting from grade and/or age differences within a single grade/age were taken into account by including a grade/age variable in each model. For examinees in Grade 4/Age 9, the grade/age variable was defined as follows:

Level	Description
1	<9 years, grade =4
2	=9 years, grade <4
3	=9 years, grade =4
4	=9 years, grade >4
5	>9 years, grade =4

Similar variables were defined for examinees in Grade 8/Age 13 and Grade 11/Age 17.

A main effects model was assumed for each grade/age. Mislevy's GROUP computer program (1984a) was used to estimate each model. In this program, examinees are grouped into a number of distinct cells based on their responses to the selected background variables. Reading proficiency, θ_i, is assumed to be normally distributed with a common variance within each cell. That is,

$$P(\theta_i \mid y_i, v_i) \sim N(v_i' \Gamma, \sigma^2)$$

where Γ is a vector of parameters corresponding to the demographic main effects and v_i is a vector characterizing the status of examinee i on those effects. Each demographic variable is represented by between one and four elements in Γ and v_i, depending on the number of levels used for that variable in the coding scheme.

The GROUP program uses an iterative procedure to estimate the elements of Γ and the common within cell variance σ^2. At each iteration, the normal distribution of reading proficiency in each cell is approximated as a histogram over 40 equally-spaced points from -4.875 to +4.875. Item parameters are assumed to be known. Sampling weights are taken into account. Iteration ends when the largest change in any effect is less than .01.

Details of the coding scheme developed for the Year 15 Reading Assessment are given in Table 10.3(2). As indicated in the table, two different methods of handling missing data were used. For some background variables, missing values were treated as valid responses, that is, a particular level of the coded variable was defined to include missing values. For example, Level 2 of the ethnicity variable, ("White and Other") includes missing values. Thus, examinees with unknown ethnicity were included in the estimation of the "White and Other" group mean. The

Table 10.3(2)

Coding of Background Variables
Year 15 BIB Data

Variable	Levels	Code	Notes
Intercept	1. All subjects	1	
Sex	1. Male	0	Subjects with missing values excluded.
	2. Female	1	
Ethnicity	1. Black	00	Subjects with missing values assigned to Level 2.
	2. White and Other	10	
	3. Hispanic	01	
STOC	1. Low Metro	00	Subjects with missing values assigned to Level 3.
	2. High Metro	10	
	3. Not High or Lo Metro	01	
Region	1. Northeast	000	Subjects with missing values excluded.
	2. Central	100	
	3. Southeast	010	
	4. West	001	
Parental Ed.	1. Less than HS	000	Subjects with missing values assigned to Level 4.
	2. High School	100	
	3. Beyond HS	010	
	4. All else	001	
Grade/Age	1. < M age, = M grade	0000	Subjects with missing values excluded. (M = modal)
	2. = M age, < M grade	1000	
	3. = M age, = M grade	0100	
	4. = M age, > M grade	0010	
	5. > M age, = M grade	0001	
Misc.	1. Subjects with unrecoverable missing data.	1	

second method of handling missing values was developed for background variables, such as sex or region, for which no single level could reasonably be defined to include missing values. Examinees with missing values for these other background variables were assigned to the "Misc." effect and were excluded from the calculations for all but the "Misc." group mean.

The dataset used to estimate conditional effects included all who responded to at least one calibrated reading item. Table 10.3(3) lists the number of examinees used to estimate each within-grade/age model along with the estimated conditional effects. An estimate of the common within-cell variance of each conditional distribution is also provided.

10.3.4 Generation of Plausible Values

A plausibility distribution was estimated for each examinee who was administered at least one of the blocks listed in Table 10.3(1). These distributions took the form of histograms over 40 equally-spaced values of reading proficiency between -4.785 and +4.785. The density of the q^{th} bar of the histogram estimated for the i^{th} examinee was obtained as follows:

$$P(A_q \mid x_i, y_i) = \frac{P(x_i \mid \theta_i = A_q, \underset{\sim}{\hat{\beta}})\, P(A_q \mid y_i, \underset{\sim}{\hat{\alpha}})}{\sum_{q=1}^{40} P(A_q \mid x_i, y_i)}$$

where

A_q = the proficiency value associated with the qth bar of the histogram, typically the midpoint of the interval;

x_i = vector of observed item responses;

y_i = responses to background and attitude items;

$\underset{\sim}{\hat{\beta}}$ = estimated item parameters;

$\underset{\sim}{\hat{\alpha}}$ = estimated conditional effects;

$P(x_i \mid \theta_i = A_q, \underset{\sim}{\hat{\beta}})$ gives the probability of observing x_i given proficiency = A_q; and

$P(A_q \mid y_i, \underset{\sim}{\hat{\alpha}})$ gives the conditional probability of A_q given background variables y_i.

Table 10.3(3)

Estimated Conditional Effects
Year 15 BIB Data

Effect	Level	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17
Intercept	All subjects	-1.350812	-0.432764	0.159135
Sex	Female	0.096410	0.139017	0.159856
Ethnicity	White and Other	0.460286	0.402945	0.405459
	Hispanic	0.076037	0.112633	0.134779
STOC	High Metro	0.490461	0.307583	0.229757
	Not High or Lo Metro	0.243873	0.122311	0.147790
Region	Central	-0.132867	-0.042057	0.027691
	South East	-0.008895	-0.020629	0.023135
	West	-0.086579	-0.042722	0.005118
Parental Ed.	High School	0.209282	0.139972	0.081576
	Beyond HS	0.395126	0.404412	0.379261
	All else	0.119694	-0.017331	-0.075156
Grade/Age	= M age, < M grade	-0.671670	-0.433070	-0.616764
	= M age, = M grade	-0.064834	-0.012745	-0.083857
	= M age, > M grade	0.338318	0.548805	0.076713
	> M age, = M grade	-0.307180	-0.259528	-0.533380
Misc.	Subjects with unrecoverable missing values.	0.509864	-0.329341	0.810939

	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17
Number of Examinees	22,950	23,553	23,932
Estimated Variances	0.46446	0.38564	0.45672

Five plausible values were obtained for each examinee by sampling at random from these histograms. For each plausible value generated, a two-step sampling procedure was required. In the first step of the procedure, a single random digit was used to target a particular block of the histogram. In the second step, a particular value of reading proficiency was chosen from within that block based on the value of a second random digit. Details of this sampling procedure are given in Table 10.3(4).

10.3.5 Effects of Specification Errors on Plausible Values

Section 10.3.1.2 discussed the possibility of biases in secondary analyses of plausible values when, during the construction of those plausible values, the true conditional distribution $p(\theta|Y)$ is approximated by some simpler approximation $p^*(\theta|Y)$. Of particular interest is the case in which $Y=(Y_1,Y_2)$, and $p^*(\theta|Y)=p(\theta|Y_1)$--that is, not all background variables are included in the conditioning process--and secondary analyses address the joint distribution of θ and Y_2. This section describes a simplified setting in which resulting biases can be derived explicitly, and employs the results to approximate the biases that would result in analyses of Year 15 NAEP reading plausible values.

The conclusion that will be reached is that secondary analyses of NAEP Year 15 reading plausible values that involve the relationship between reading proficiency and non-conditioned variables (e.g., subgroup means, regression analyses, and path analyses) must be interpreted with caution, because the strength of these relationships will tend to be underestimated by amounts that depend on the type of analysis and the inter-relationships of the variables involved. Numerical results for selected analyses are presented below in Section 10.3.5.4. The strength of relationships between reading proficiency and conditioned variables only, on the other hand, will not be underestimated. Comparisons of regression coefficients or multiple correlations with reading proficiency may thus prove misleading, to a degree whose magnitude is suggested by a number of examples from the reading database.

The remainder of this section provides the foundation upon which this conclusion is based.

10.3.5.1 Setup and Notation

As mentioned in Section 10.3.1.2, closed-form expressions for secondary biases under the IRT model used in the NAEP reading analysis are not readily forthcoming. We therefore derive results for a related but simpler context, namely the classical true-score measurement model:

$$X = \theta + \varepsilon,$$

Table 10.3(4)

Sampling Procedure Used to Generate Plausible Values

Step 1: Obtain a random number r from the unit interval. Select bar k from the histogram estimated for the i^{th} examinee such that

$$\sum_{q=1}^{k-1} P(A_q \mid x_i, y_i) < r \leq \sum_{q=k}^{40} P(A_q \mid x_i, y_i)$$

where

$P(A_q \mid x_i, y_i)$ = Density of the q^{th} bar, with value A_q, for q=1,...40

x_i = vector of observed item responses, and

y_i = vector of responses to background and attitude items.

Step 2: Obtain a second random number s from the unit interval. Compute the plausible value θ_i as follows:

$$\theta_i = A_k + .2(s - .5)$$

where

θ is the unobservable variable of interest,

X is the observable variable, and

ε is a random error variable.

The following distributional assumptions will be made:

$$\varepsilon \approx N(0,\ \sigma_{\varepsilon}^{2}) \qquad\qquad \mathrm{Cov}(\varepsilon,\theta)=0$$

$$\theta \approx N(\mu,\ \sigma_{\theta}^{2});$$

it follows that $X \approx N(\mu,\ \sigma_{x}^{2})$ with $\sigma_{x}^{2} = \sigma_{\theta}^{2} + \sigma_{\varepsilon}^{2}$, and $\mathrm{Cov}(\theta,X)= \sigma_{\theta}^{2}$.

The following normal linear regression model is assumed for the examinee population:

$$\theta = \beta' Y + F,$$

where Y is a K-dimensional vector of background variables, with $Y \approx MVN(0,\Sigma)$ and $F \approx N(0,\sigma_{\theta|Y}^{2})$ where $\sigma_{\theta|Y}^{2} = \sigma_{\theta}^{2} - \beta'\Sigma\beta$. Note that

$$\underline{E}(\theta|Y) = \underline{E}(X|Y) = \beta' Y, \text{ and}$$

$$\mathrm{Var}\ (X|Y) = \sigma_{\varepsilon}^{2} + \sigma_{\theta|Y}^{2} \ .$$

Define the "conditional" reliability ρ of X given Y as follows:

$$\rho = \sigma_{\theta|Y}^{2} / [\sigma_{\theta|Y}^{2} + \sigma_{\varepsilon}^{2}].$$

Note that $0<\rho<1$ and $\rho\sigma_{X|Y}^{2} = \sigma_{\theta|Y}^{2}$.

In a generalization of Kelley's (1947) formula (see also Box & Tiao, 1973, p. 74), we find that

$$\underline{E}(\theta|X,\ Y) = \rho X + (1-\rho)\ \beta' Y \text{ and}$$

$$\mathrm{Var}\ (\theta|X,\ Y) = (1-\rho)\ \sigma_{\theta|Y}^{2} \ .$$

That is, the expected value of the unobservable variable of interest, θ, given the values of observable variables X and Y, is a weighted average of (i) the imperfect manifestation X, in the proportion that it is "reliable," and (ii) the expected value of the unobservable variable given background information Y, in the proportion that X is unreliable. Note that $\theta|X,Y$ follows a normal distribution under our simplifying assumptions.

10.3.5.2 Plausible Values, Complete Conditioning

Under the preceding assumptions, a plausible value $\tilde{\theta}$ is obtained in the following manner:

$$\begin{aligned} \tilde{\theta} &= \tilde{\theta}\ (X,\ Y) \\ &= \underline{E}(\theta|X,\ Y) + G \\ &= \rho X + (1 - \rho)\ \beta' Y + G, \end{aligned}$$

where G is a random number selected from $N(0,(1-\rho)\sigma^2_{\theta|Y})$, independently of X and Y. The following properties of $\tilde{\theta}$ are derived in Mislevy (in progress):

$$\begin{aligned} &(1)\ \underline{E}(\tilde{\theta}|Y = y) && = \underline{E}(\theta|Y = y) && = \beta' y \\ &(2)\ \underline{E}(\tilde{\theta}) && = \underline{E}(\theta) && = \mu \\ &(3)\ \mathrm{Var}\ (\tilde{\theta}|Y = y) && = \mathrm{Var}\ (\theta|Y = y) && = \sigma^2_{\theta|Y} \\ &(4)\ \mathrm{Var}\ (\tilde{\theta}) && = \mathrm{Var}\ (\theta) && = \sigma^2_{\theta} \end{aligned}$$

These results indicate that the expected value of the analyses listed above, when carried out with plausible values and combinations of observable variables--is identical to the expected value when carried out with θ itself--an intrinsically unobservable variable. Even though the θ values of specific individuals remain unknown, and θ values may in fact serve poorly as estimates for individual respondents, the method of constructing the plausible values yields the correct results for population characteristics.

10.3.5.3 Plausible Values, Incomplete Conditioning

Suppose that Y can be partitioned into two subvectors, Y_1 and Y_2. The same population structure holds, though it may be rewritten to reflect the partitioning as

$$\underline{E}(\theta) = \beta'_1\ Y_1 + \beta'_2\ Y_2$$

and

$$\Sigma = \begin{matrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{matrix}$$

Define the projection operators $P_1 = \Sigma_{21}\Sigma_{22}^{-1}$ and $P_2 = \Sigma_{12}\Sigma_{11}^{-1}$. P_1 and P_2 possess the following properties:

(1) $\underline{E}(Y_2 | Y_1 = y_1) = P_2 y_1$ and $\underline{E}(Y_1 | Y_2 = y_2) = P_1 y_2$.

(2) The j^{th} diagonal element of $P_2 P_1$ is the squared multiple correlation of the j^{th} element in Y_2 with the elements of Y_1.

(3) Regression coefficients for θ and X on Y_1 alone or on Y_2 alone can be expressed as follows:

$$\underline{E}(\theta | Y_1 = y_1) = \underline{E}(X | Y_1 = y_1) = \beta_1^{*\prime} y_1 \equiv (\beta_1' + \beta_2' P_2)\, y_1$$

and

$$\underline{E}(\theta | Y_2 = y_2) = \underline{E}(X | Y_2 = y_2) = \beta_2^{*\prime} Y_2 \equiv (\beta_2' + \beta_1' P_1)\, y_2 \ .$$

(4) If Y_1 and Y_2 are orthogonal, both P_1 and P_2 are matrices of zeros.

(5) Intuitively, $P_1 P_2 Y_1$ yields the portion of Y_1 predictable by Y_2; a similar relationship holds for the relationship of $P_2 P_1 Y_2$ to Y_1.

Suppose now that plausible values $\tilde{\theta}^*$ are constructed that take into account the relationship of θ with Y_1, but not with Y_2. That is,

$$\begin{aligned} \tilde{\theta}^* &= \tilde{\theta}^* \quad (X, Y_1, Y_2) \\ &= \underline{E}(\theta | X_1 Y_1) + G^* \\ &= \rho^* X + (1 - \rho^*)\, \beta_1^{*\prime} Y_1 + G^*, \end{aligned}$$

where

$$\rho^* = \frac{\sigma^2_{\theta | Y_1}}{\sigma^2_{\theta | Y_1} + \sigma^2_E}$$

and G^* is selected at random from $N(0,(1-\rho^*)\sigma^2_{\theta|Y_1})$, independently of X and Y_1.

It follows immediately from the results of the preceding subsection that

$$(1)\ \underline{E}(\tilde{\theta}^*|Y_1 = y_1) = \underline{E}(\theta|Y_1 = y_1) = \beta_1^{*\prime} y_1$$

$$(2)\ \underline{E}(\tilde{\theta}^*) = \underline{E}(\theta) = \mu$$

$$(3)\ \mathrm{Var}\ (\tilde{\theta}^*|Y_1 = y_1) = \mathrm{Var}\ (\theta|Y_1 = y_1) = \sigma^2_{\theta|Y_1}$$

$$(4)\ \mathrm{Var}\ (\tilde{\theta}^*) = \mathrm{Var}\ (\theta) = \sigma^2_\theta \quad .$$

These analyses involving properties of the distribution of θ in the population at large, and of its relationship with Y_1, have the same expected value when carried out with $\tilde{\theta}^*$ as with θ.

Analyses involving Y_2 do not fare as well, however. Key results, again derived in Mislevy (in progress), are summarized below.

(1) Whereas $\underline{E}(\theta|Y_1 = y_1, Y_2 = y_2) = \beta_1' y_1 + \beta_2' y_2$,

we find that

$$\underline{E}(\tilde{\theta}^*|Y_1 = y_1, Y_2 = y_2) = \beta_1' y_1 + (1-\rho^*)\beta_2' P_2 y_1 + \rho^*\beta_2' y_2 \tag{1a}$$

$$= \beta_1' y_1 + (1-\rho^*)\ \beta_2'\ \underline{E}(Y_2|Y_1 = y_1) + \rho^*\beta_2' y_2 \tag{1b}$$

$$= \beta_1' y_2 + \beta_2' y_2 - (1-\rho^*)\ \beta_2'\ (y_2 - P_2 y_1) \tag{1c}$$

A bias thus results in the construction of a plausible value for a respondent with values of y_1 and y_2 on the background variables. The contribution from Y_1 is correct, but the contribution from Y_2 is attenuated. Rather than a contribution from that person's y_2 value, we obtain a weighted average of the contribution from his or her particular y_2 to the extent that X is reliable, but from the expected value of Y_2 given his or her particular y_1 to the extent that X is unreliable. It follows from (1c) that this bias can be driven to zero in three different ways:

(i) If $\rho^* = 1$; i.e., X is a perfectly reliable measure of θ;

(ii) If $\beta_2 = 0$; i.e., there is no contribution from Y_2 anyway;

(iii) If $P_2 y_1 = y_2$; i.e., if $\underline{E}(Y_2 | Y_1 = y_1) = y_2$. This will be true for all y_2 only if Y_2 is perfectly predictable from Y_1.

Bias in the expected value of the plausible values of individual subjects is mitigated as any or all of these conditions are approached.

(2) Whereas $\underline{E}(\theta | Y_2 = y_2) = (\beta_2' + \beta_1' P_1) y_2 = \beta_2^{*\prime} \, y_2$, we find

$$\underline{E}(\tilde{\theta}^* | Y_2 = y_2) = \{\beta_2' \, [\rho^* + (1-\rho^*) P_2 P_1] + \beta_1' \, P_1\} y_2 \tag{2a}$$

$$= [\beta_2^{*\prime} - \beta_2' \, (1-\rho^*)(I - P_2 P_1)] y_2 \quad . \tag{2b}$$

As in (1) above, it can be seen in (2a) that the contribution relating to the Y_1 space comes through correctly, but the contribution of the Y_2 space is again the average of the actual y_2 (to the extent that X is reliable) and just the portion of y_2 that is predictable through Y_1 (to the extent that X is unreliable). The bias term is reduced as either ρ^* or the proportion of Y_2 predictable from Y_1 approaches 1.

(3) Whereas the regression coefficient for y_2 in the multiple regression of θ on Y_1 and Y_2 can be found through

$$\underline{E}[\theta - \underline{E}(\theta | Y_1) \;|\; Y_2 = y_2] = \beta_2' \, y_2 ,$$

we find that

$$\underline{E}[\tilde{\theta}^* - \underline{E}(\tilde{\theta}^* \, | Y_1) | Y_2 = y_2] = \beta_2' \, \rho^* \, (I - P_2 P_1) \, y_2 \; .$$

Compared with the desired result, namely β_2', we must expect a shrunken answer when we run the regression with $\tilde{\theta}^*$. Shrinkage is mitigated as ρ^* approaches 1 and as $P_2 P_1$ approaches zero.

(4) Whereas the regression coefficient for y_1 in the multiple regression of θ on Y_1 and Y_2 can be found through

$$\underline{E}[\theta - \underline{E}(\theta | Y_2) | Y_1 = y_1] = \beta_1' y_1 ,$$

we find that

$$\underline{E}[\tilde{\theta}^* - \underline{E}(\tilde{\theta}^* | Y_2) | Y_1 = y_1] = [\beta_1' + (1-\rho^*) \beta_2' \, P_2](I - P_1 P_2) y_1 .$$

Thus bias appears in the multiple regression coefficient for Y_1, even though it has been conditioned upon, unless $\rho^*=1$ and $P_1P_2=0$.

Two aspects of these results have sobering implications for secondary analyses of plausible values. First, while higher reliability ρ^* is unequivocally helpful, high shared variance between Y_1 and Y_2 is not. High shared variance mitigates shrinkage when the simple regression of θ on Y_2 is approximated by the regression of $\tilde{\theta}^*$ on Y_2; high shared variance exacerbates shrinkage with respect to the coefficient for Y_2 when the multiple correlation of θ on Y_1 and Y_2 is approximated by the same analysis of $\tilde{\theta}^*$.

Second, a particularly popular form of secondary analysis is threatened when both conditioned and non-conditioned background variables are involved. Specifically, the size of the simple regression coefficient of proficiency on a given background variable is often compared with the corresponding partial regression coefficient when other predictors are included in the model. A decrease in the size of the coefficient of the focus variable is expected, presumably heading from β^* toward β as more explanatory variables are taken into account. Result (4) above indicates that if the focus variable was conditioned upon while the other explanatory variables were not, then the partial regression coefficient for the focus variable will contain a contribution properly associated with the other variables. In the extreme case of $\beta_1=0$, in particular, the expected result will not be 0.

10.3.5.4 Approximating Secondary Biases in the Analysis of Year 15 NAEP Reading Plausible Values

The analyses above assume normality, linearity, knowledge of population covariance matrices, and an observable variable X that is related to θ in the same manner for all respondents. None of these assumptions is strictly satisfied in the NAEP database. They may prove useful nonetheless by illustrating the order of magnitude of the secondary biases that will exist in secondary regression analyses of NAEP reading plausible values. This section describes how one may compute approximate "unshrunken" coefficients for non-conditioned variables taken one at a time, both in simple regression analyses and in multiple regression analyses with the entire set of conditioning variables. The steps are as follows:

(1) Calculate an approximate ρ^*, or $\hat{\rho}^*$. An approximation computed from number-right scores within NAEP booklets can be averaged over booklets. This value will tend to underestimate the precision of the IRT analyses, since number-right scores capture most, but not all, of the information available in response patterns. The steps to be carried out in each booklet are as follows:

(a) Calculate a reliability coefficient, say Cronbach's alpha, for percent-correct scores X (with omits and not-reached treated as incorrect). This yields

$$\hat{\rho} = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_E} \equiv \frac{\sigma^2_{\theta|Y_1} + \sigma^2_{\hat{\theta}}}{\sigma^2_{\theta|Y_1} + \sigma^2_{\hat{\theta}} + \sigma^2_E},$$

where $\sigma^2_{\hat{\theta}}$ is the variance of $\varepsilon(\theta|Y_1)$, with Y_1 being the NAEP conditioning variables.

(b) Compute the proportion of variance in X accounted for by Y_1 by standard ANOVA procedures:

$$\hat{R}^2_\theta = \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\hat{\theta}} + \sigma^2_{\theta|Y} + \sigma^2_E} \equiv \frac{\sigma^2_{\hat{x}}}{\sigma^2_x}$$

since $\sigma^2_{\hat{\theta}} = \sigma^2_{\hat{x}}$

(c) Because $\rho^* \equiv \sigma^2_{\theta|Y_1} / (\sigma^2_{\theta|Y_1} + \sigma^2_E)$,

$$\hat{\rho}^* = \frac{\hat{\rho} - \hat{R}^2_\theta}{1 - \hat{R}^2_\theta}$$

Values of $\hat{\rho}^*$ as calculated from each of the NAEP booklets are shown in Tables 10.3(5), 10.3(6) and 10.3(7).

Average per-booklet reliability coefficients (rho) are .82, .75, and .77 for Grade/Ages 4/9, 8/13, and 11/17 respectively; corresponding average proportions of variance of percent-correct scores explained by conditioning variables are .26, .25, and .28; and reliabilities after partialling out conditioning variables (rho-star) are .75, .67, and .68. Recall that it is these latter values that set the tone for the magnitude of shrinkage effects to be expected in secondary analyses of non-conditioned variables. A fairly strong relationship will be noted

Table 10.3(5)

Reliability Coefficients by Booklet
Grade 4/Age 9

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
1	7	0.79	0.25	0.72
2	20	0.87	0.29	0.82
5	10	0.75	0.19	0.69
6	10	0.76	0.18	0.71
7	35	0.92	0.37	0.87
8	24	0.83	0.21	0.79
9	30	0.90	0.37	0.84
11	12	0.71	0.23	0.63
12	10	0.77	0.21	0.71
13	23	0.81	0.27	0.75
14	21	0.87	0.24	0.83
15	12	0.65	0.18	0.57
16	18	0.84	0.19	0.80
17	22	0.87	0.28	0.83
18	13	0.82	0.30	0.74
19	18	0.82	0.28	0.76
20	12	0.79	0.21	0.74
21	12	0.80	0.31	0.72
22	32	0.90	0.34	0.85
23	17	0.86	0.28	0.80
24	12	0.75	0.21	0.68
25	28	0.87	0.27	0.83
26	11	0.76	0.25	0.68
27	20	0.81	0.31	0.73
28	20	0.78	0.22	0.71
29	22	0.83	0.31	0.75
30	20	0.87	0.34	0.81
31	21	0.82	0.26	0.76
32	19	0.84	0.40	0.74
33	23	0.83	0.33	0.74
34	11	0.66	0.25	0.54
35	23	0.87	0.25	0.82
36	11	0.55	0.27	0.38
37	11	0.70	0.21	0.63
38	22	0.80	0.27	0.73
39	13	0.82	0.31	0.74
40	7	0.78	0.23	0.71
41	31	0.88	0.33	0.83
42	9	0.72	0.18	0.65
43	9	0.76	0.21	0.70
45	24	0.88	0.30	0.84
46	12	0.84	0.22	0.79

Table 10.3(5)
(continued)

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
47	24	0.88	0.29	0.83
48	12	0.80	0.26	0.73
49	36	0.87	0.27	0.82
50	25	0.85	0.24	0.80
51	23	0.87	0.36	0.80
53	35	0.92	0.31	0.89
54	11	0.74	0.23	0.66
55	24	0.88	0.25	0.84
56	23	0.89	0.30	0.84
57	22	0.87	0.32	0.81
58	18	0.84	0.24	0.79
59	20	0.87	0.25	0.82
60	21	0.85	0.27	0.79
61	10	0.75	0.19	0.70
63	11	0.80	0.23	0.75
MEAN	18.46	0.82	0.26	0.75

Table 10.3(6)

Reliability Coefficients by Booklet
Grade 8/Age 13

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
1	6	0.56	0.20	0.45
2	15	0.79	0.37	0.66
5	12	0.68	0.17	0.62
6	12	0.66	0.16	0.60
7	31	0.89	0.20	0.86
8	22	0.85	0.27	0.80
9	28	0.86	0.26	0.82
11	11	0.67	0.20	0.58
12	12	0.71	0.15	0.65
13	19	0.79	0.22	0.73
14	22	0.84	0.35	0.76
15	11	0.62	0.27	0.48
16	23	0.79	0.28	0.70
17	23	0.83	0.27	0.77
18	9	0.69	0.24	0.60
19	14	0.76	0.27	0.67
20	12	0.81	0.30	0.73
21	12	0.78	0.26	0.70
22	37	0.86	0.33	0.80
23	18	0.75	0.19	0.70
24	11	0.75	0.20	0.69
25	27	0.85	0.28	0.79
26	17	0.70	0.19	0.62
27	18	0.80	0.23	0.74
28	20	0.76	0.32	0.65
29	27	0.77	0.30	0.68
30	27	0.82	0.34	0.72
31	22	0.75	0.25	0.66
32	17	0.76	0.22	0.70
33	21	0.83	0.29	0.76
34	10	0.50	0.25	0.34
35	28	0.82	0.35	0.72
36	10	0.51	0.26	0.34
37	8	0.67	0.25	0.55
38	18	0.72	0.27	0.62
39	9	0.59	0.20	0.49
40	6	0.50	0.15	0.41
41	33	0.86	0.24	0.82
42	10	0.64	0.26	0.51
43	10	0.67	0.29	0.53
45	17	0.79	0.27	0.71
46	12	0.79	0.29	0.71

Table 10.3(6)
(continued)

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
47	26	0.81	0.33	0.72
48	11	0.77	0.20	0.71
49	30	0.80	0.25	0.74
50	20	0.81	0.27	0.74
51	28	0.80	0.31	0.70
53	33	0.87	0.33	0.81
54	8	0.69	0.18	0.63
55	23	0.85	0.27	0.79
56	29	0.82	0.29	0.74
57	19	0.75	0.31	0.64
58	22	0.81	0.25	0.74
59	18	0.85	0.27	0.79
60	11	0.74	0.15	0.69
63	6	0.73	0.17	0.68
MEAN	18.05	0.75	0.25	0.67

Table 10.3(7)

Reliability Coefficients by Booklet
Grade 11/Age 17

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
1	6	0.61	0.20	0.51
2	16	0.80	0.25	0.73
5	12	0.75	0.31	0.64
6	12	0.75	0.20	0.69
7	27	0.90	0.35	0.84
8	18	0.82	0.34	0.73
9	31	0.85	0.35	0.77
11	11	0.71	0.27	0.60
12	12	0.72	0.35	0.57
13	19	0.82	0.33	0.72
14	17	0.85	0.34	0.77
15	11	0.66	0.21	0.58
16	17	0.70	0.31	0.57
17	19	0.83	0.37	0.74
18	10	0.78	0.25	0.70
19	14	0.78	0.27	0.70
20	12	0.80	0.31	0.71
21	12	0.78	0.32	0.67
22	31	0.86	0.28	0.81
23	18	0.73	0.25	0.64
24	7	0.72	0.23	0.63
25	18	0.80	0.29	0.71
26	11	0.68	0.25	0.57
27	13	0.77	0.24	0.70
28	18	0.77	0.29	0.68
29	24	0.81	0.40	0.68
30	16	0.70	0.28	0.59
31	25	0.83	0.39	0.72
32	17	0.76	0.26	0.67
33	20	0.85	0.30	0.79
34	13	0.71	0.26	0.60
35	22	0.82	0.29	0.75
36	13	0.74	0.29	0.64
37	8	0.67	0.29	0.53
38	21	0.78	0.36	0.65
39	10	0.70	0.25	0.61
40	6	0.62	0.13	0.56
41	28	0.85	0.30	0.79
42	5	0.41	0.15	0.30
43	5	0.54	0.19	0.43
45	18	0.81	0.19	0.76
46	12	0.82	0.30	0.74

Table 10.3(7)
(continued)

BOOKLET	# ITEMS	RHO	R-2	RHO-STAR
47	21	0.78	0.33	0.67
48	7	0.72	0.19	0.66
49	34	0.89	0.32	0.84
50	17	0.80	0.24	0.74
51	18	0.80	0.35	0.69
53	34	0.90	0.37	0.85
54	8	0.78	0.27	0.70
55	23	0.84	0.30	0.77
56	23	0.80	0.39	0.68
57	15	0.79	0.26	0.72
58	17	0.80	0.20	0.75
59	18	0.85	0.26	0.80
60	7	0.74	0.23	0.67
63	6	0.79	0.10	0.77
MEAN	16.13	0.77	0.28	0.68

between the numbers of items in booklets and their reliability coefficients.

(2) For a given background variable Y_2 (or a single contrast involving Y_2 if it is a categorical variable with several categories), compute the squared multiple correlation R^2_Y between Y_2 and Y_1.

(3) Run the multiple regression analysis for plausible values that includes Y_1 and Y_2 as predictors. The expected coefficient for Y_2 is $B_2 = \beta'_2 \rho^* (1-R^2_Y)$. From this, one can calculate an "unshrunken" estimate of the partial regression coefficient:

$$\hat{\beta}_2 = \hat{B}_2 / \hat{\rho}^* (1- \hat{R}^2_Y).$$

(4) Run the simple regression of $\tilde{\theta}^*$ on Y_2. The expected coefficient is $B'^*_2 = \beta^{*\prime}_2 - \beta'_2 (1-\rho^*)(1-R^2_Y)$. From this, and the result of Step 3, one can calculate an "unshrunken" estimate of the simple regression coefficient for Y_2:

$$\hat{\beta}^*_2 = \hat{B}^*_2 + \hat{\beta}'_2 (1-\hat{\rho}^*)(1-\hat{R}^2_Y).$$

Tables 10.3(8) through 10.3(10) carry out these steps for a number of non-conditioned NAEP variables in the three grade/age samples. The column labeled "R-square" indicates the proportion of variance in the focal non-conditioned variable that is accounted for by the conditioning variables; that is, R^2_Y. The columns headed "Multiple regression" concern the estimation of a regression coefficient for the focal variable in a multiple regression equation containing it and all conditioning variables as predictors of proficiency. The columns headed "Simple regression" concern the simple regression of proficiency upon the focal variable alone.

Results for the simple regressions indicate shrinkages from about 5 to 15 percent, with the average about 10 percent. As expected, shrinkages of this type are smallest when the focal variable exhibits comparatively higher shared variation with the conditioning variables. The percentage of pupils in a respondent's school that participate in federal lunch programs is an example. This variable is related to ethnicity, parents' education, and type of community, resulting in an R-square of about .2. Simple regression shrinkage is thus minimal for this variable--about 5 percent.

Results for the multiple regressions indicate shrinkages between 25 and 45 percent, with the average about 35 percent. Now, variables with high shared variation exhibit greater shrinkages. The shrinkage for "percent in lunch program" is 38 percent for the youngest subsample (where reliability is highest) and about 45 percent for the two older subsamples.

To repeat the introduction to this section, we arrive at two conclusions. First, secondary analyses involving relationships between NAEP Year 15 reading plausible values and conditioning variables provide,

Table 10.3(8)

Approximate Shrinkage of Regression Coefficients of Non-conditioned Background Variables: Grade 4/Age 9

		Multiple Regression			Simple Regression		
Effect	R-square*	B	Beta	% Shrink	B-star	Beta-star	% Shrink
Hours TV	.04	-.79	-1.10	27.78	-1.63	-1.89	13.93
Papers read	.03	4.92	6.75	27.03	10.18	11.81	13.81
% pupils in lunch pgm	.18	-.07	-.11	38.03	-.23	-.25	8.96
Minority school	.11	-1.20	-1.80	33.28	-9.51	-9.90	4.01
School problems	.07	1.93	2.78	30.36	5.28	5.92	10.82
Minority reading pgm	.05	-11.92	-16.73	28.77	-34.03	37.98	10.40
Classes taken	.03	.54	.74	27.30	2.14	2.32	7.67

* Proportion of variation of focal effect accounted for by conditioning variables

Table 10.3(9)

Approximate Shrinkage of Regression Coefficients of Non-conditioned Background Variables: Grade 8/Age 13

		Multiple Regression			Simple Regression		
Effect	R-square*	B	Beta	% Shrink	B-star	Beta-star	% Shrink
Hours TV	.03	.11	.17	34.89	.33	.39	14.22
Papers read	.03	6.26	9.65	35.07	12.33	15.41	19.96
% pupils in lunch pgm	.20	-.02	-.04	46.25	-.19	-.20	5.94
Minority school	.15	1.63	2.86	43.04	-3.88	-3.08	-25.92
School problems	.07	1.40	2.25	37.81	3.26	3.94	17.46
Minority reading pgm	.07	-3.72	-5.97	37.69	-27.14	-28.97	6.30
Classes taken	.08	1.75	2.84	38.42	3.69	4.55	18.87

* Proportion of variation of focal effect accounted for by conditioning variables

Table 10.3(10)

Approximate Shrinkage of Regression Coefficients of Non-conditioned Background Variables: Grade 11/Age 17

		Multiple Regression			Simple Regression		
Effect	R-square*	B	Beta	% Shrink	B-star	Beta-star	% Shrink
Hours TV	.06	-1.06	-.65	36.03	-3.20	-3.70	13.42
Papers read	.02	9.80	14.73	33.49	15.03	19.63	23.43
% pupils in lunch pgm	.20	-.03	-.06	45.28	-.30	-.31	4.59
Minority school	.14	1.55	2.66	41.69	-7.99	-7.26	-10.03
School problems	.12	1.17	1.94	39.81	5.26	5.80	9.44
Minority reading pgm	.01	-6.45	-9.60	32.81	-33.70	-36.73	8.25
Academic courses	.18	1.10	1.98	44.19	2.80	3.32	15.62

* Proportion of variation of focal effect accounted for by conditioning variables

by construction, consistent estimates of modeled effects. Second, in the way of contrast, analyses that involve the relationship of reading proficiency and non-conditioned variables must be interpreted with caution, because the strength of these relationships will tend to be underestimated. The underestimation is least serious when only the margin of a single unconditioned variable is addressed (about 10 percent on the average), but more serious when higher order features of the joint relationship of proficiency, conditioning variables, and unconditioned variables is addressed, as in multiple regression (35 percent on the average).

10.3.6 BIB/Pace Equating

In addition to the reading responses solicited under BIB administration conditions during the Year 15 assessment, responses were solicited from national probability samples of age-eligible pupils under the pace conditions that characterized past NAEP assessments. These pace data play a pivotal role in NAEP. Because they were obtained under the same administration conditions as data from past assessments, they make possible the continuation of unbroken trend lines from the past. Because they were obtained from samples of pupils that were randomly equivalent to those of the age-based BIB samples, they make possible the linkage of BIB and pace results. This section describes the steps that were taken to link the Year 15 pace results into the Year 15 BIB scale discussed above. Chapter 10.4 will describe the analysis of trends, from past assessments through the Year 15 pace assessment.

10.3.6.1 Percent-Correct Results

Before discussing the IRT procedures employed to equate the spiralled and paced data, it is useful to examine descriptive characteristics of the data of the two types in terms of percentages of correct responses. In this more familiar metric, the reader may more easily judge for himself or herself the magnitude of the effect of administration, its consistency over items, and the degree of differential effects among gender and ethnicity groups.

Figures 10.3-3 through 10.3-8 plot weighted percents-correct from spiralled administration against those of paced administration for the same item among Age 9 whites, blacks, Hispanics, males, and females. Only those items included in IRT scaling are shown; plots incorporating non-IRT items are also available from ETS upon request. Figures 10.3-9 through 10.3-14 present similar results for Age 13, and Figures 10.3-15 through 10.3-20 present results for Age 17.

Inspection of these plots reveals only modest administration effects, mainly linear and similar for most items. Due to the greater precision of their percent-correct values, plots involving the larger groups (male, female, and white) exhibit less scatter than plots for smaller groups (black and Hispanic). More detailed comparisons can be obtained from Table 10.3(11), which presents correlations among percents-correct under the two

Figure 10.3-3

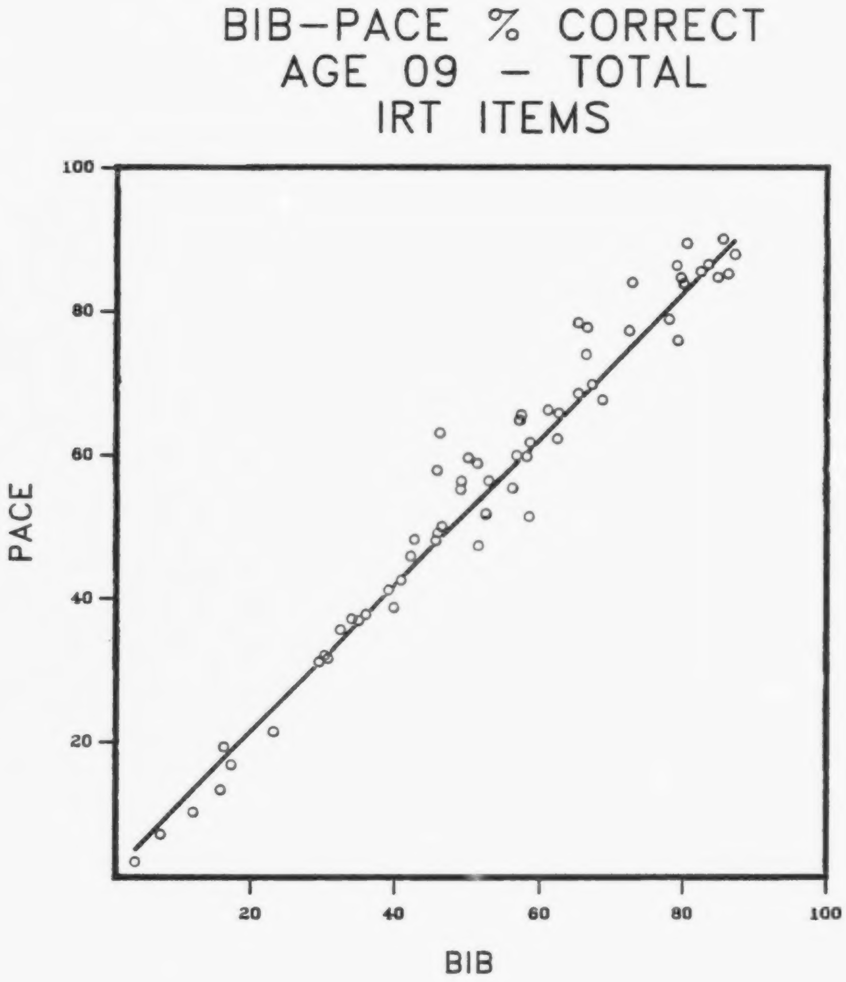

BIB–PACE % CORRECT
AGE 09 – TOTAL
IRT ITEMS
PACE
BIB
100
80
60
40
20
20
40
60
80
100

Figure 10.3-4

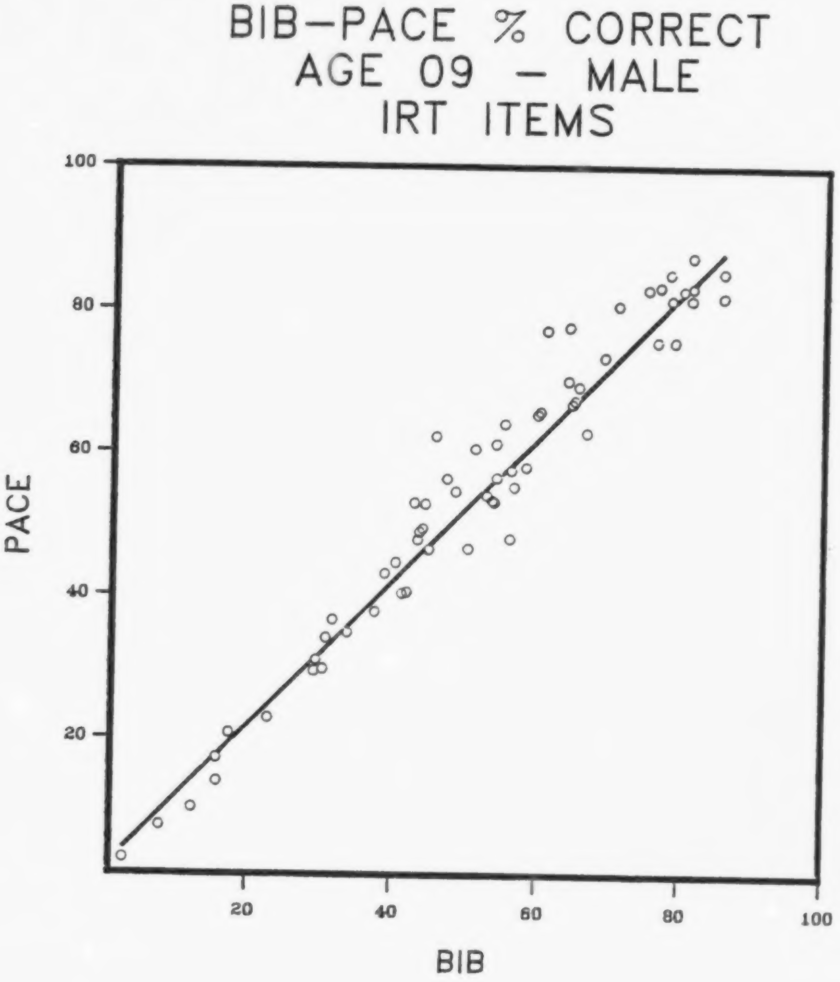

BIB-PACE % CORRECT
AGE 09 - MALE
IRT ITEMS
PACE
BIB
100
80
60
40
20
20
40
60
80
100

Figure 10.3-5

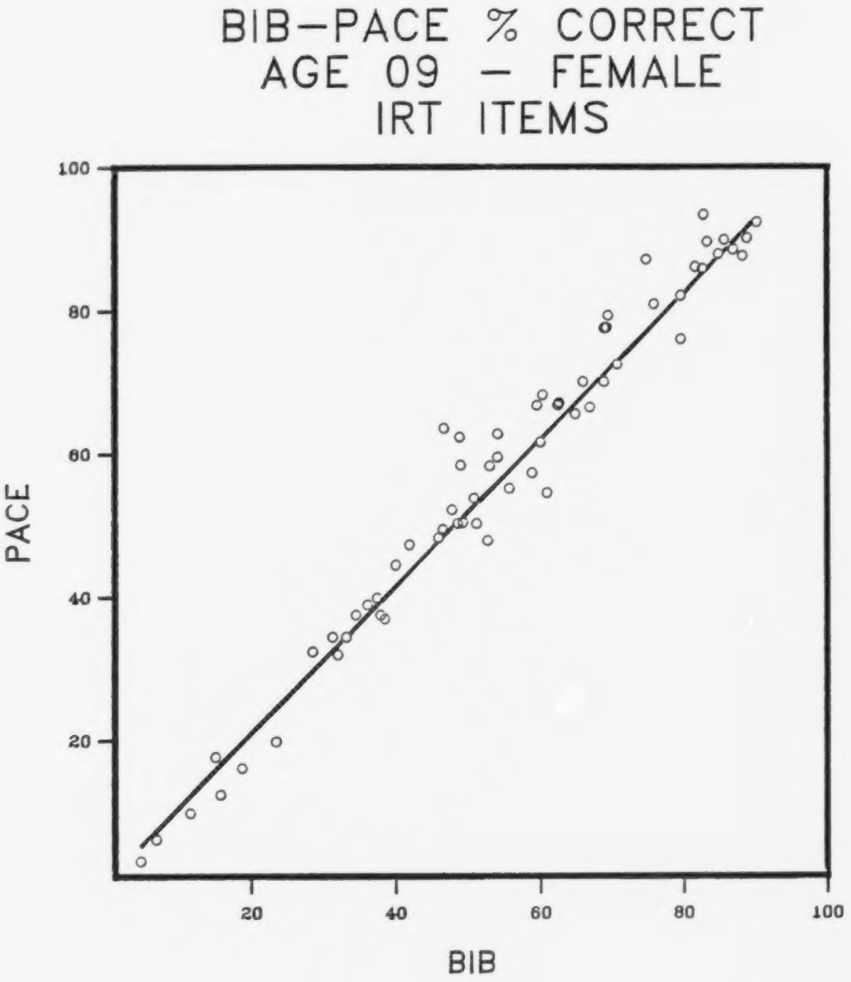

BIB-PACE % CORRECT
AGE 09 - FEMALE
IRT ITEMS
PACE
BIB
100
80
60
40
20
20
40
60
80
100

Figure 10.3-6

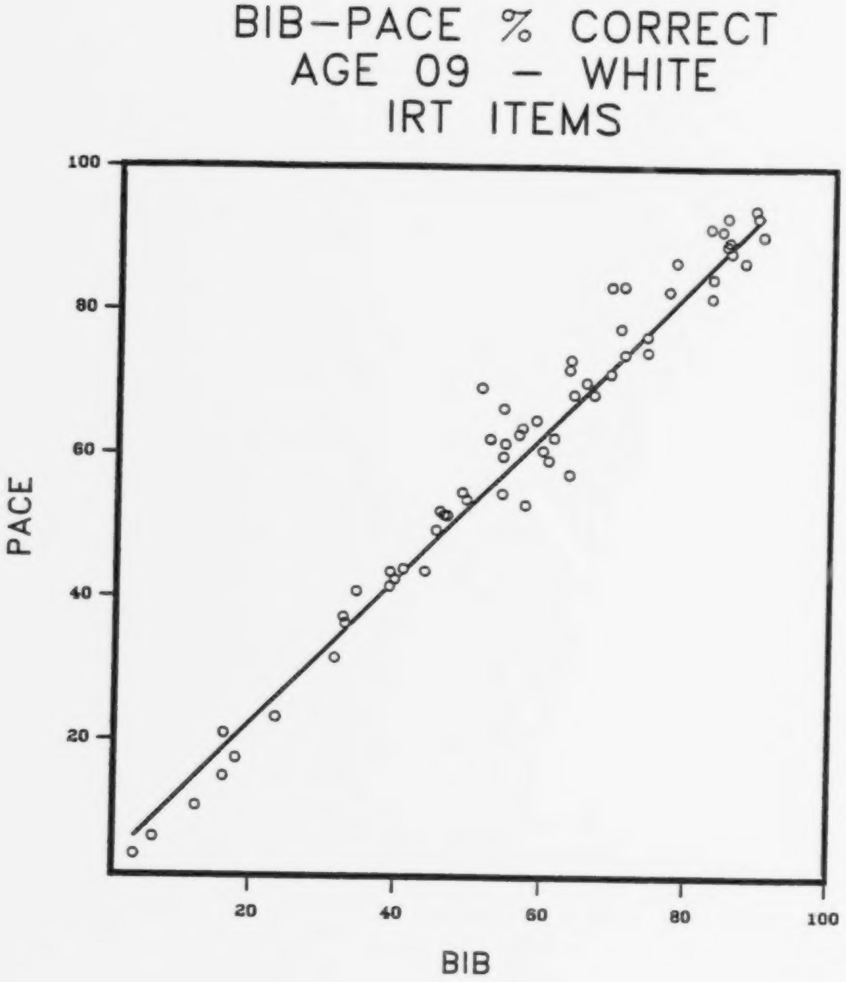

BIB-PACE % CORRECT
AGE 09 - WHITE
IRT ITEMS
100
80
60
40
20
PACE
20
40
60
80
100
BIB

Figure 10.3-7

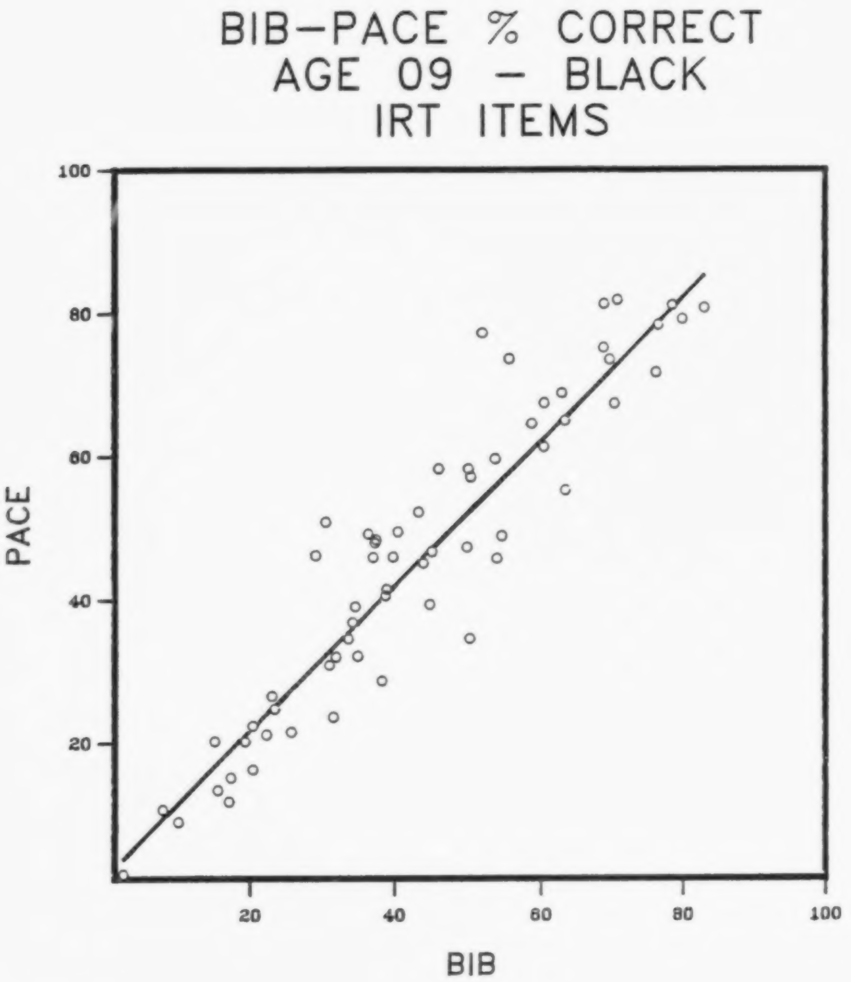

BIB-PACE % CORRECT
AGE 09 - BLACK
IRT ITEMS
100
80
60
40
20
PACE
20
40
60
80
100
BIB

Figure 10.3-8

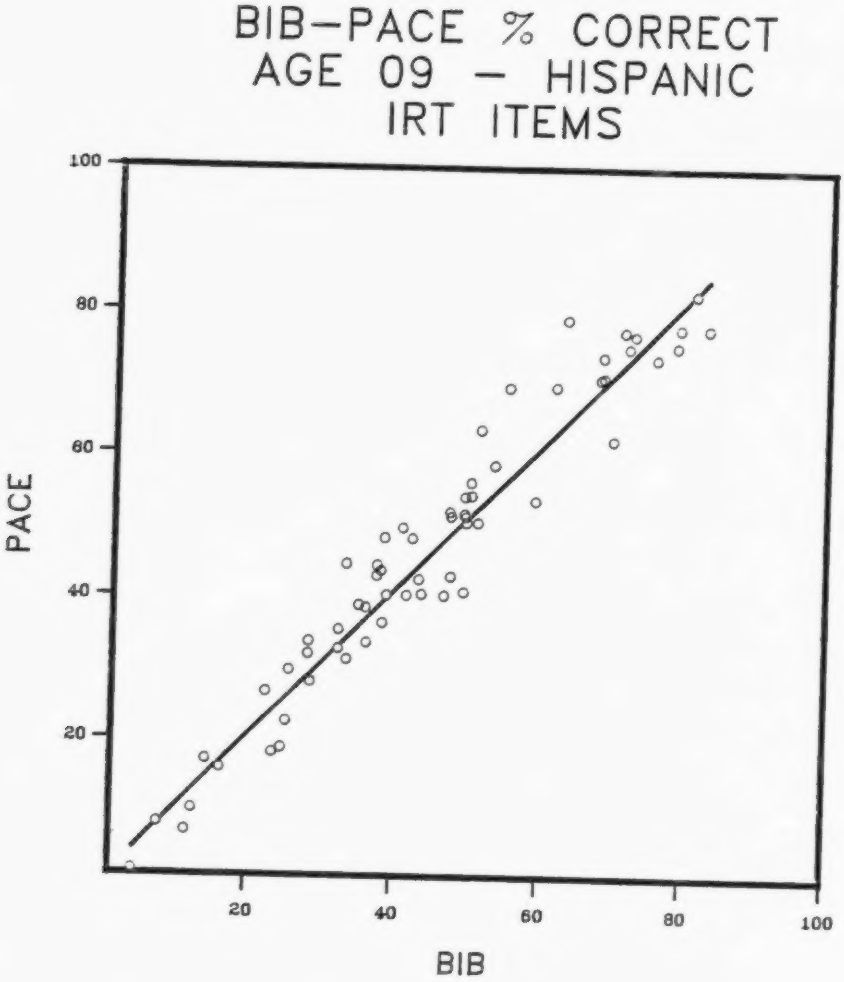

BIB-PACE % CORRECT
AGE 09 - HISPANIC
IRT ITEMS
PACE
BIB
100
80
60
40
20
20
40
60
80
100

Figure 10.3-9

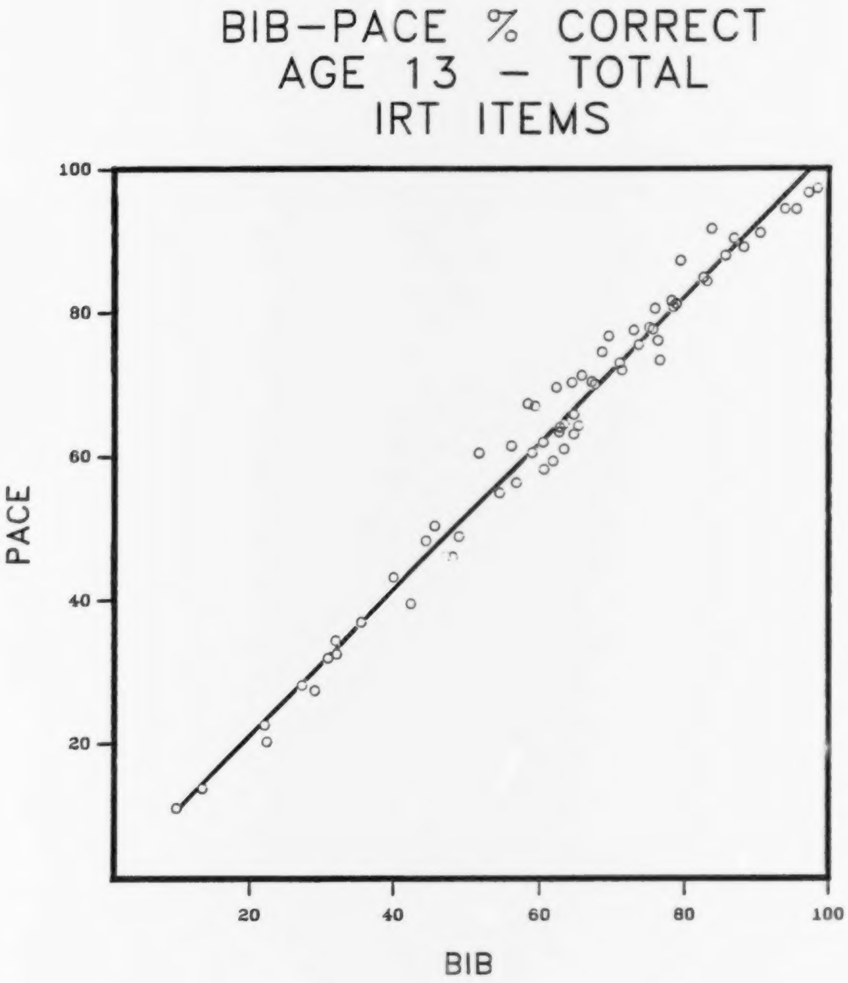

BIB-PACE % CORRECT
AGE 13 - TOTAL
IRT ITEMS
PACE
BIB
100
80
60
40
20
20
40
60
80
100

Figure 10.3-10

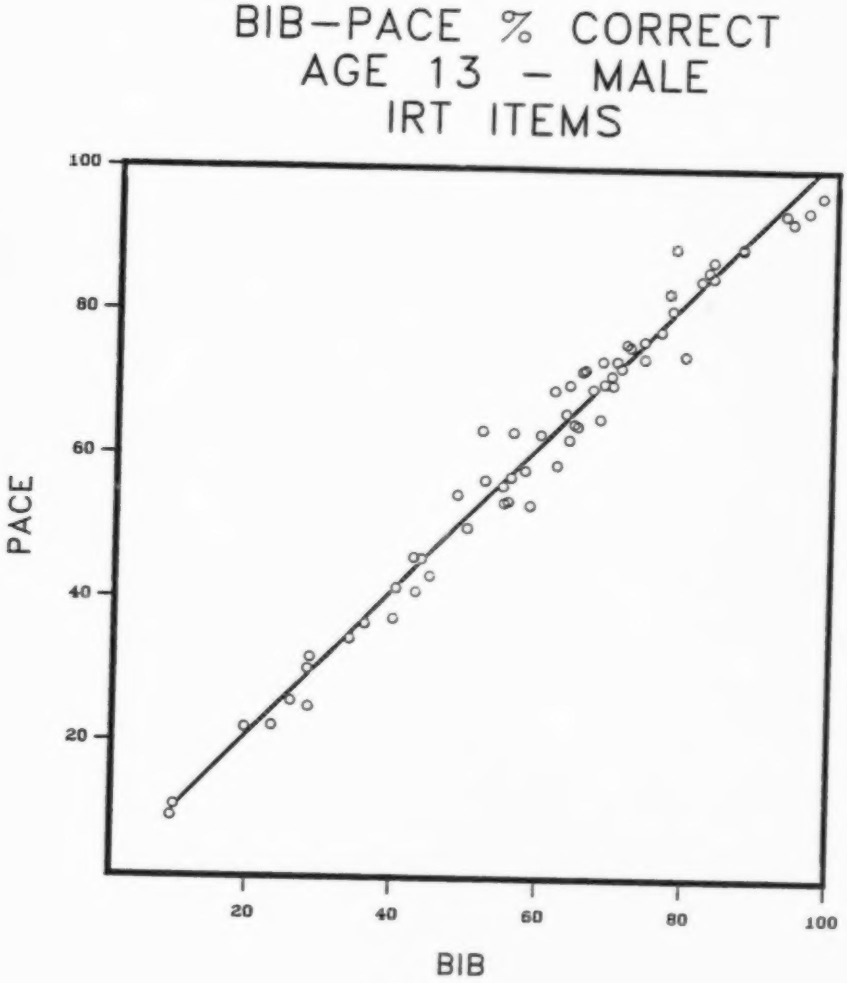


Figure 10.3-11

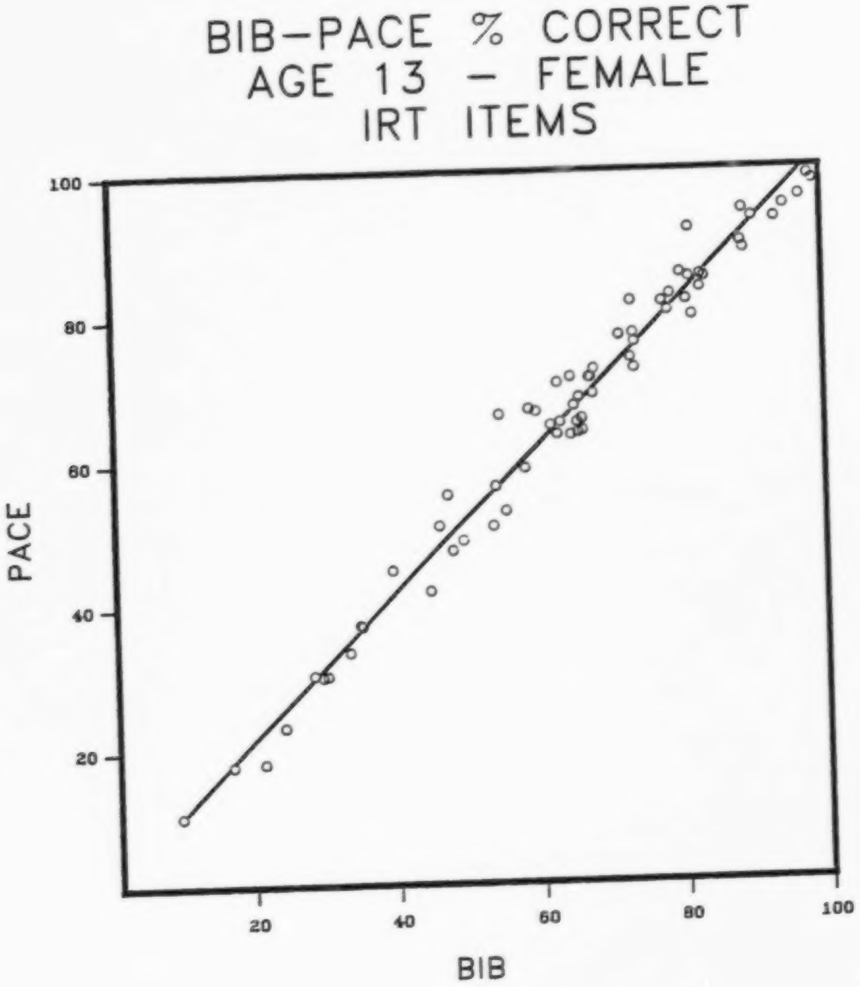


Figure 10.3-12

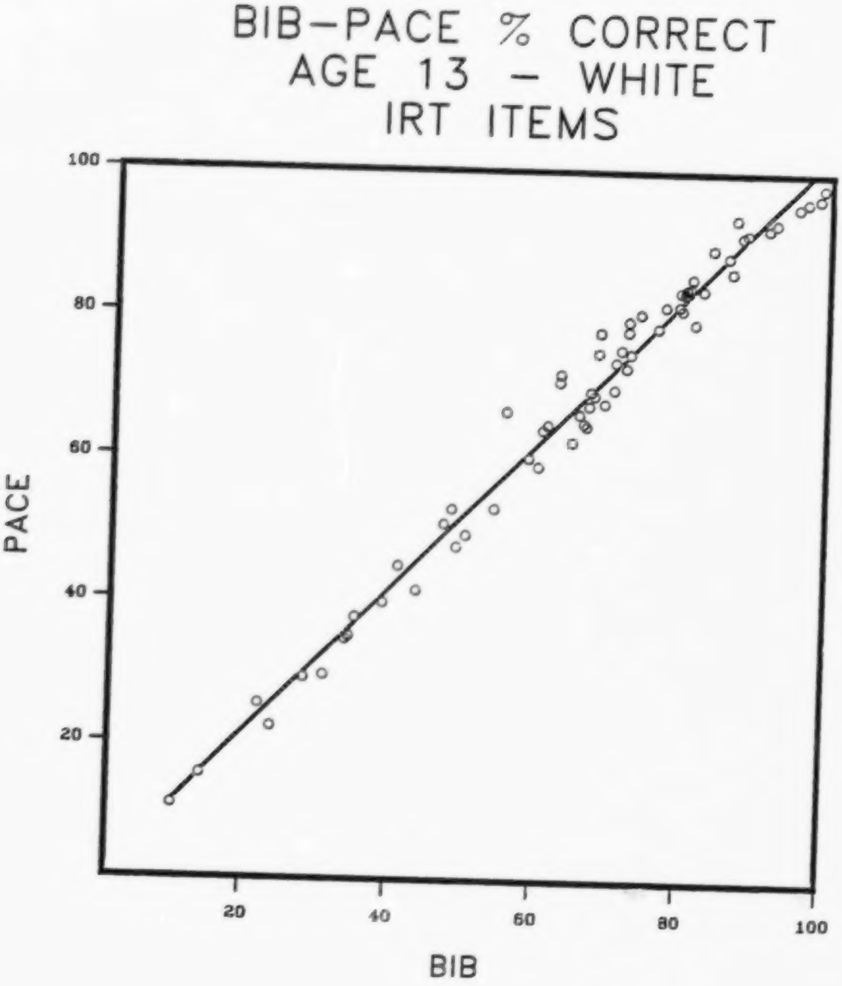


Figure 10.3-13

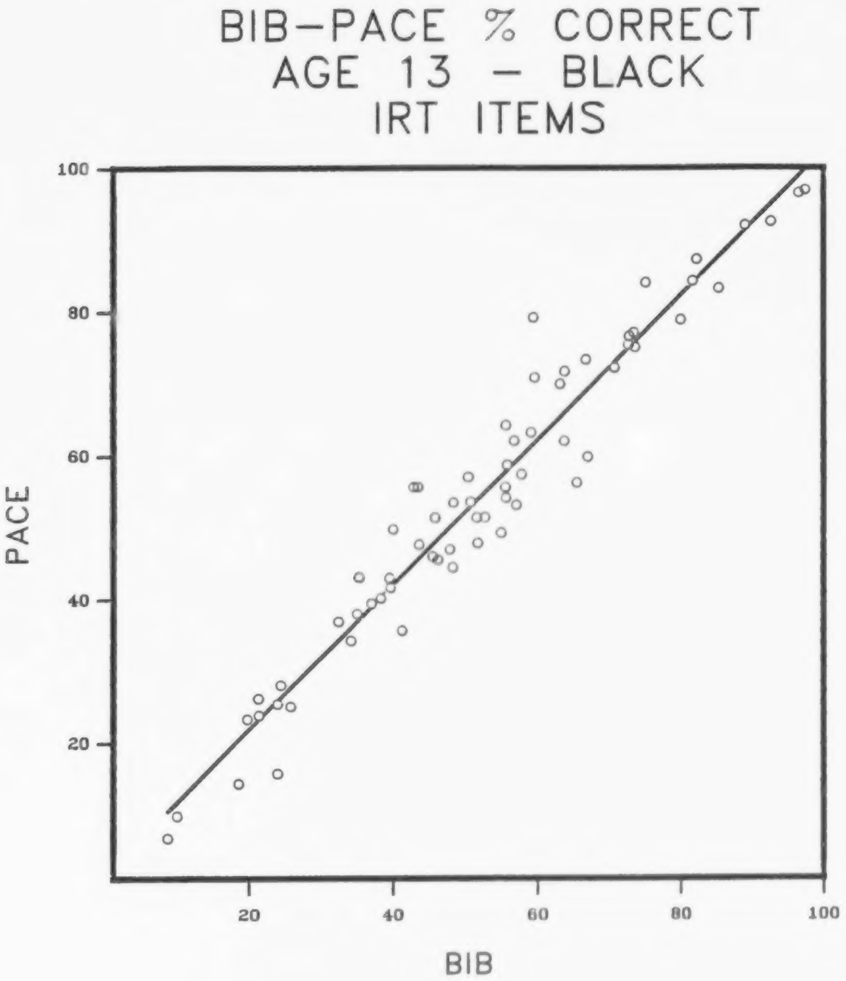


Figure 10.3-14

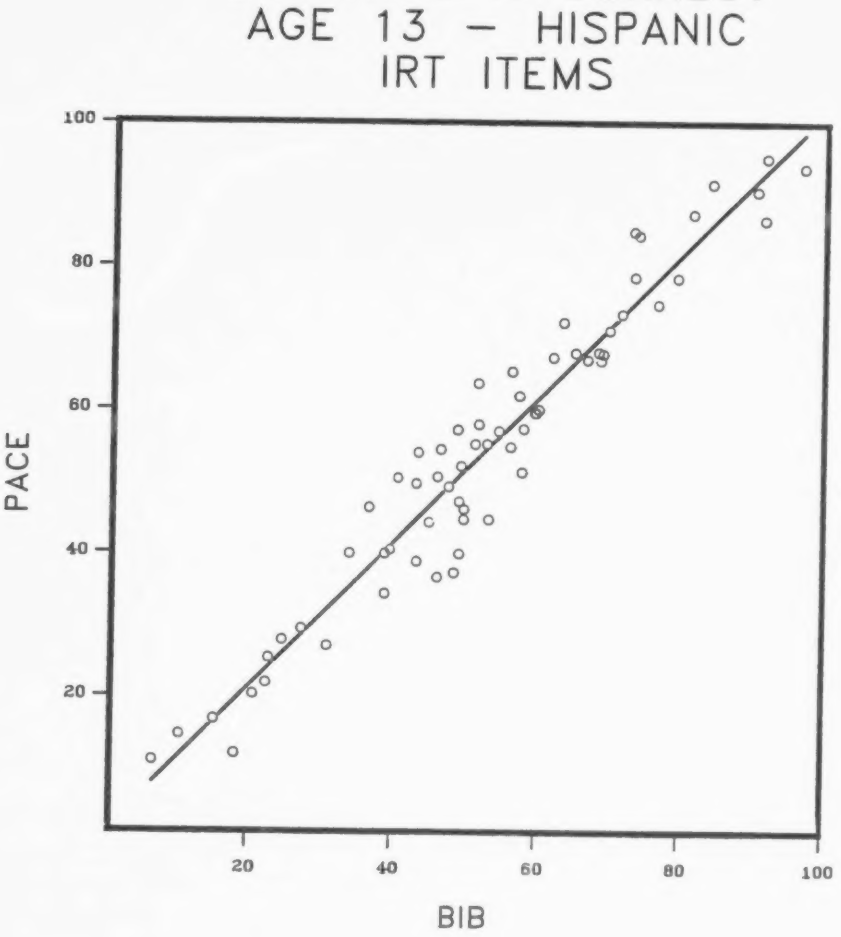


Figure 10.3-15

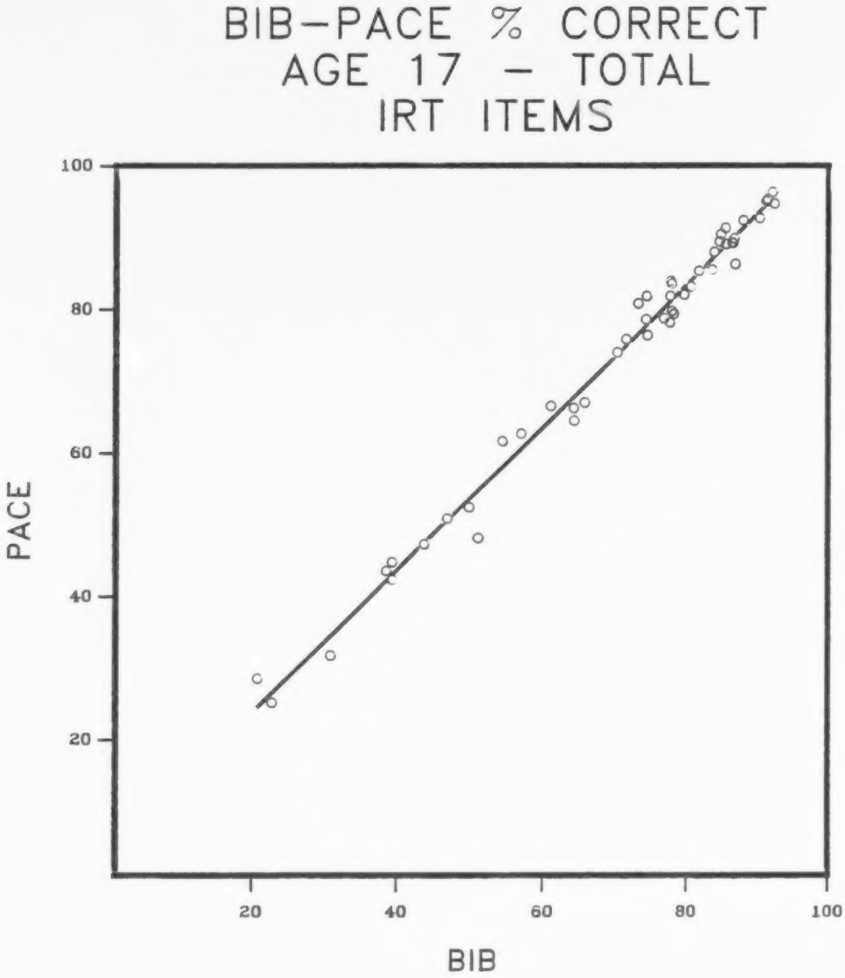


Figure 10.3-16

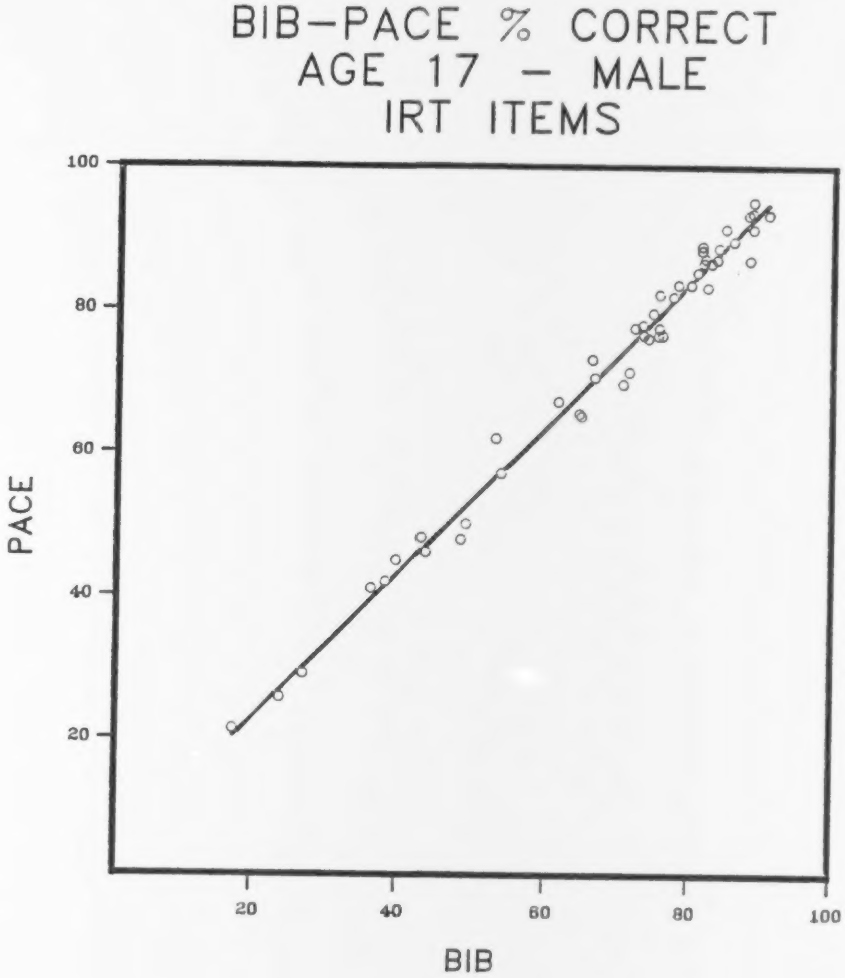


Figure 10.3-17

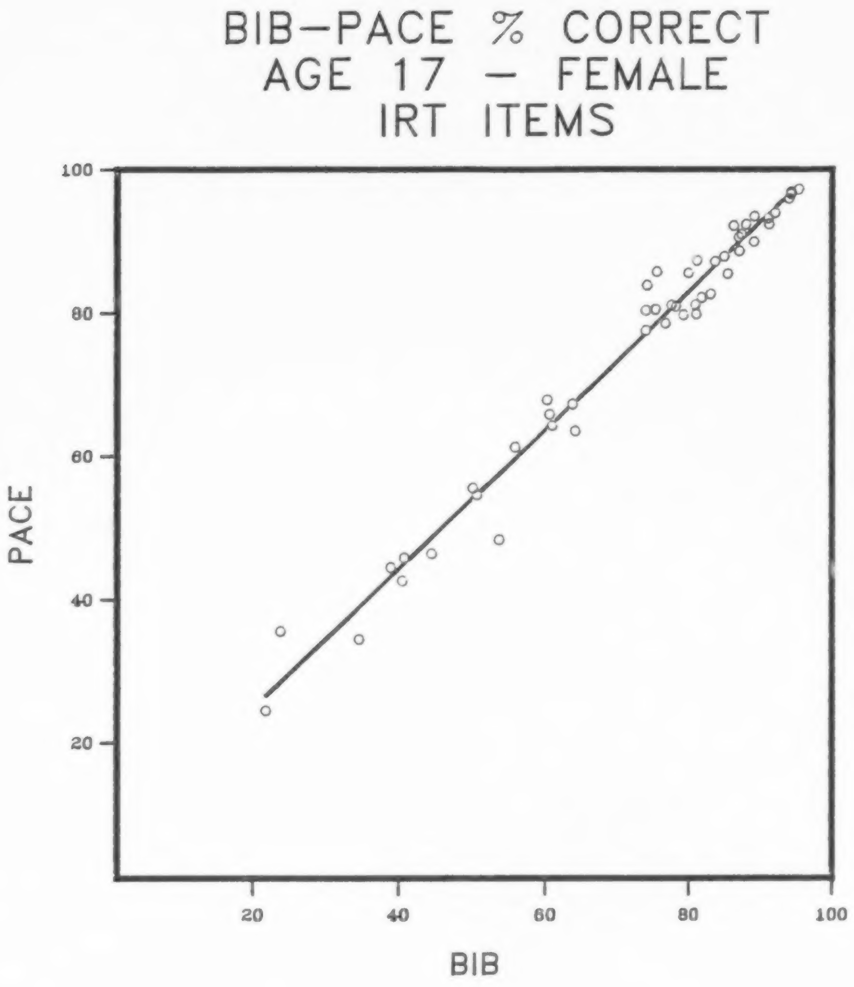


Figure 10.3-18

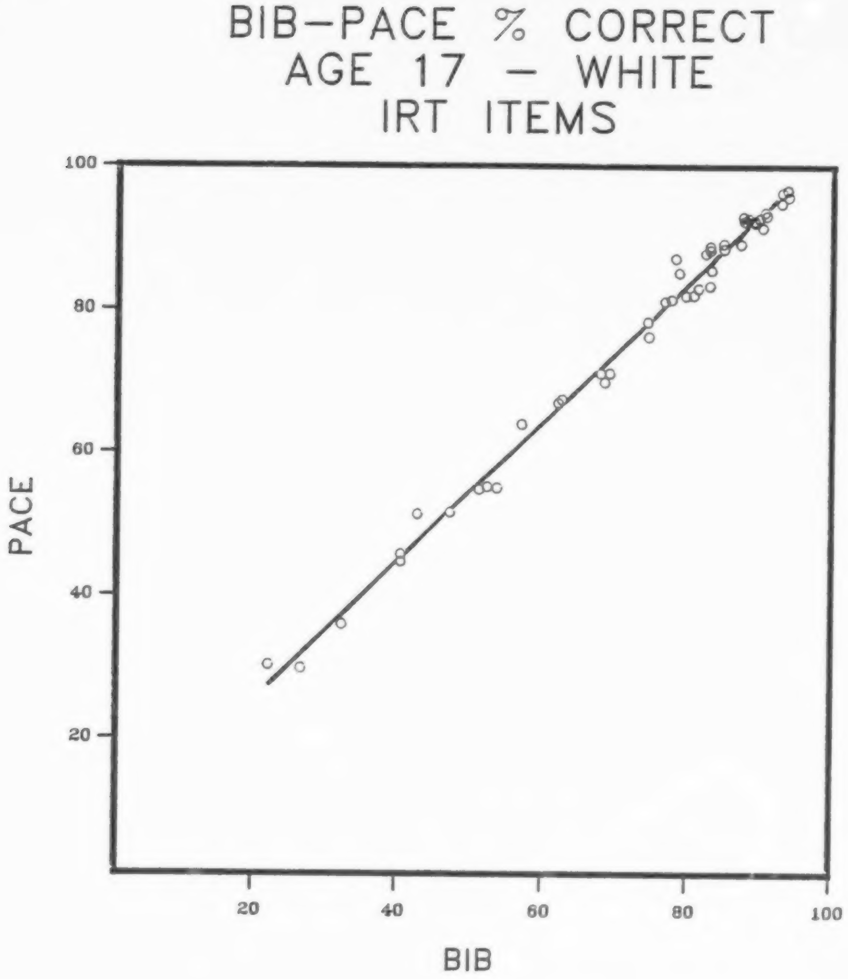


Figure 10.3-19

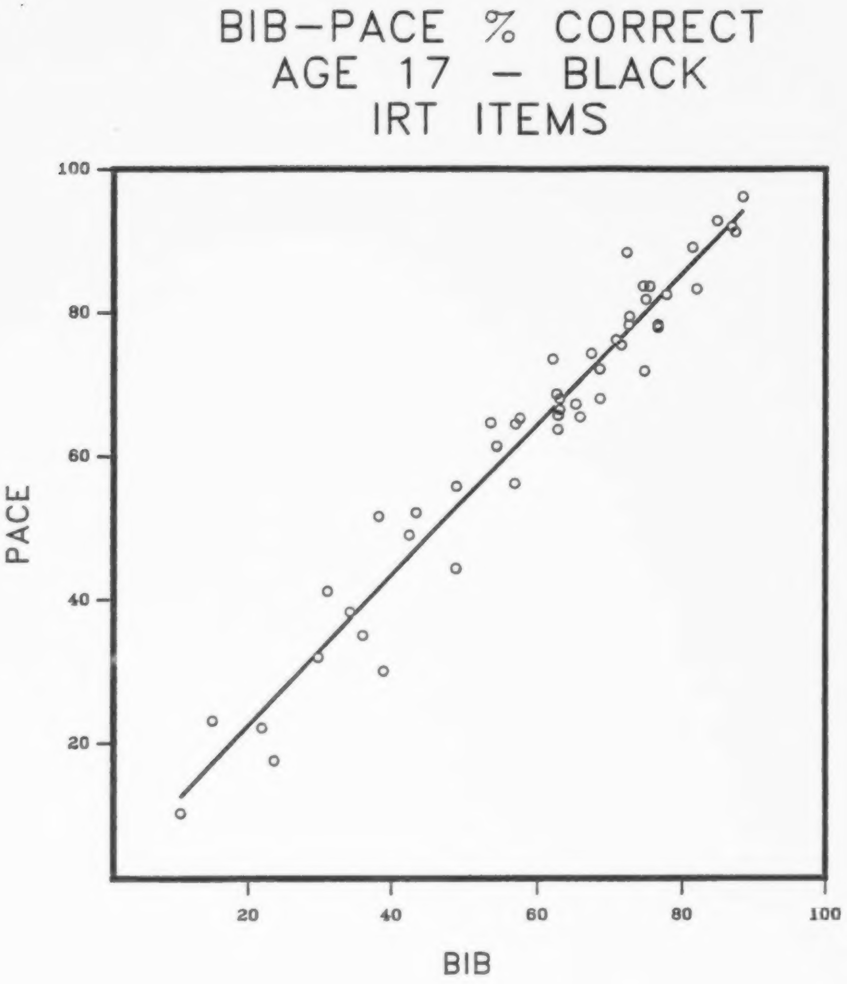


Figure 10.3-20

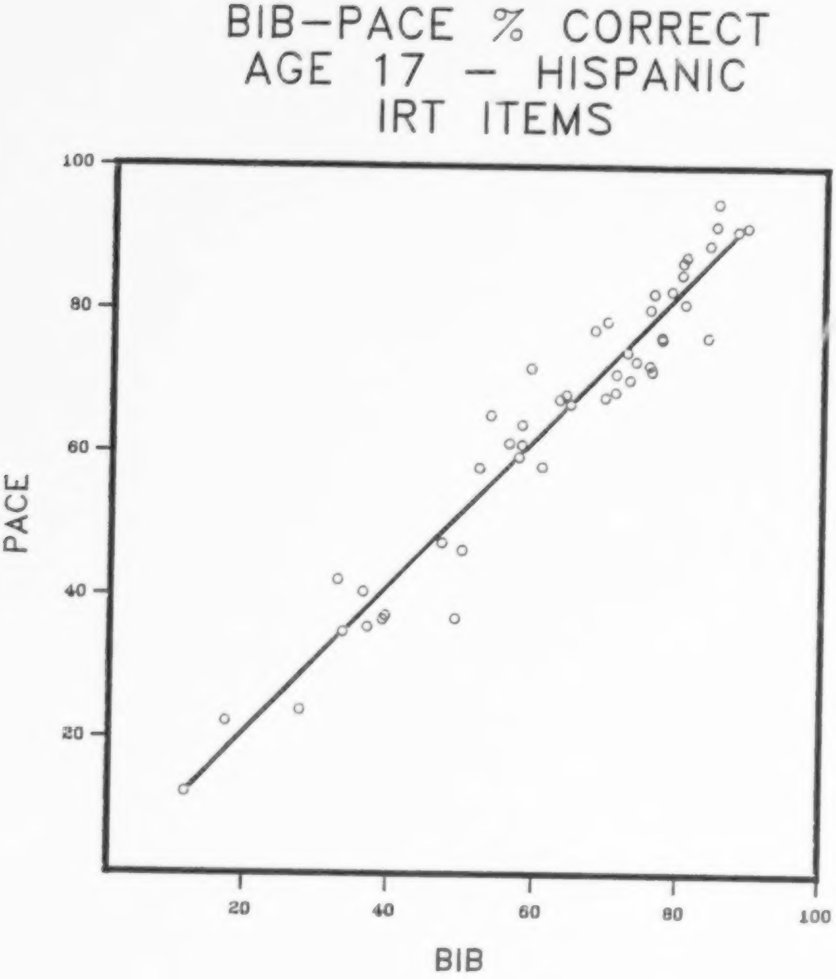


Table 10.3(11)

Correlations and Regression Coefficients for Spiral vs. Pace Percents-Correct of IRT Reading Items

Group		Correlation	Regression Coefficients*	
			Intercept (se)	Slope (se)
Age 9	Total	.91	1.05 (3.31)	1.02 (.06)
	Male	.91	1.66 (3.28)	1.01 (.06)
	Female	.92	.73 (3.39)	1.02 (.06)
	White	.91	2.35 (3.63)	1.01 (.06)
	Black	.89	1.60 (3.10)	1.01 (.06)
	Hispanic	.92	.22 (2.65)	1.02 (.05)
Age 13	Total	.99	.67 (1.19)	1.01 (.02)
	Male	.99	.68 (1.31)	1.02 (.02)
	Female	.99	1.07 (1.31)	1.01 (.02)
	White	.99	1.23 (1.24)	1.02 (.02)
	Black	.97	1.74 (1.75)	1.01 (.03)
	Hispanic	.97	.91 (1.91)	1.01 (.03)
Age 17	Total	.99	3.79 (1.21)	1.00 (.02)
	Male	.99	2.46 (1.28)	1.02 (.02)
	Female	.99	5.54 (1.67)	.97 (.02)
	White	.99	5.80 (1.05)	.97 (.01)
	Black	.97	1.46 (2.22)	1.05 (.04)
	Hispanic	.97	.26 (2.53)	1.03 (.04)

*Regression of paced percent-correct on spiralled percent-correct, in percentage-point units.

administrations, and regression lines predicting paced percent-corrects from spiralled percents-correct. The main results are as follows.

(1) While the effect of mode of administration varies somewhat from item to item, the average effect was for items to be slightly easier under paced conditions than BIB. As indicated by the intercept coefficient, the size of the effect was on the order of 1 percentage point in Ages 9 and 13, and was not statistically significant; it was about 4 percentage points for age seventeen, and was statistically significant.

(2) Item-by-administration interactions were negligible at Ages 13 and 17 (as evidenced by correlations of .99), but were somewhat more apparent at Age 9 (a correlation of .91). Patterns of interactions, as seen in the plots, are similar across gender and ethnicity groups.

(3) No significant gender-by-administration or ethnicity-by-administration interactions were observed in Ages 9 or 13. At Age 17, however, the effect of pacing favored females over males (5.5 percentages points to 2.5), and whites over blacks and Hispanics (5.8 percentages points to 1.5 and .3).

This last finding suggests that the change from paced tape administration had little effect, if any, on the relative differences in performance among subgroups. In particular, the new procedures do not appear to have had a detrimental effect on the performance of black and Hispanic students.

10.3.6.2 IRT Equating Procedures

The three-parameter logistic IRT model was employed for the Year 15 BIB data and for pace data (Year 15 pace, plus all past reading assessments). For reasons to be made clear below, the pace scaling was carried out separately within each age. There is no guarantee that the reading proficiency variables defined in these separate analyses are in fact measuring exactly the same skill. In fact, there is reason to suspect they are not, since under BIB conditions the pupils had to read directions and control the amount of time they spent on each item for themselves.

The practical question is whether a reasonably straightforward relationship can be maintained between the variables, such that the same family of IRT models and the same reporting scale can be used for BIB data and pace data from all ages and years. If the same IRT model is to be used, the form of the three-parameter logistic model restricts the relationship among the variables measured under BIB and under pace at the three ages to be linear. That is, if the probability that person i from age group k will answer item j correctly under BIB conditions is given by

$$P(X_{ij} = 1 | a_j, b_j, c_j, \theta_i) = c_j + (1-c_j) \Psi [1.7a_j(\theta_i - b_j)] ,$$

then the probability of a correct response under pace conditions must be of the form

$$P_k^\star (X_{ij} = 1 | a_j, b_j, c_j, \theta_i) = c_j + (1-c_j)\Psi\{1.7(a_j/M_k)[\theta_i - (M_k b_j + X_k)]\} ,$$

where M_k and X_k are constants that apply to all items in the domain.

Allowing different constants M and X for different age groups allows for the possibility that BIB/pace differences may interact with age. To the degree that such interactions exist, pace results from different ages translated to the BIB scale may not be strictly comparable. Comparisons within ages are comparable over all time points, however, as are comparisons within age from BIB to pace.

The assumption of a linear relationship on the θ scale between BIB and pace data can be estimated by use of the randomly equivalent BIB and pace age samples at each age level. One way that this can be accomplished is to fit the three-parameter model to the BIB and pace data separately (as was in fact done), and translate the pace scale (from its arbitrary origin and unit-size) in a manner that matches the first two moments of the age distribution to that obtained from the BIB sample.

To this end, the model was fit to the Year 15 BIB data in the manner described in the preceding sections, and to all pace data (Year 15 pace plus all past assessments) for each age separately in the manner described in the following section. The distributions of the Year 15 BIB samples in each age (on the Year 15 reading proficiency θ scale), and of each Year 15 pace age samples (on a provisional scale with an arbitrary origin and unit-size) were estimated by means of the computer program RESOLVE (Mislevy, 1985c), which provided a non-parametric approximation of the latent θ distribution in terms of a histogram. It should be noted that this procedure estimates the latent distribution directly rather than using point estimates for individuals, thereby avoiding difficulties associated with infinite estimates for aberrant response patterns and with differing precision of tests with different lengths (see Mislevy, 1984b, for details).

It is important to note that the assumption of a linear relationship can be checked in three ways. Matching the first two moments of the distributions ensures that these two characteristics of the BIB and pace data will agree; other characteristics have not been constrained to match, but should match fairly well if the assumption is reasonable. After matching moments in the manner described above, three checks of the veracity of the linearity assumption were performed:

(1) Match of distributional shape. Distributions with identical means and variances can differ considerably with respect to

features such as skewness, kurtosis, and multiplicity of modes. Such findings would indicate that (at best) a curvilinear rather than linear relationship would be required to match BIB and pace θ's, thus precluding the possibility of using a joint three-parameter logistic model for responses gathered under both conditions. The plots obtained from the three age-sample comparisons are shown as Figures 10.3-21 through 10.3-23. They indicate a reasonably good match of higher-order features.

(2) Match of subgroup means. While the means and standard deviations from an age sample as a whole were constrained to match, means of population subgroups were not. Matches at the level of subgroup means are crucial, however, if comparability is to be claimed; even if the distributions for the population as a whole were quite similar, the finding that differences among major subgroups reversed orders or shifted dramatically in magnitude would also preclude the pooling of BIB and pace results. Tables 10.3(12) through 10.3(14) present population and subgroup means from the three ages, as computed from BIB and pace data after the first two population moments have been matched. Consistent agreement as to order and relative magnitudes of differences among major subgroups are uniformly evident.

(3) Match of item parameters. The parameter estimates of items taken by any pair of ages under pace conditions and estimated separately within age-group data, should be in essential agreement after the scales of the separate pace age-group analyses are translated to the Year 15 BIB scale. Pairwise plots for the estimates of b parameters for items administered to Ages 9 and 13, and 13 and 17, appear as Figures 10.3-24 and 10.3-25. Excellent agreement is shown for the 13 versus 17 plot; good agreement is shown for the 9 versus 13 plot, except that items that were difficult for both ages (high b values) tended to be relatively even more difficult for 9-year-olds under pace conditions.

Taken together, these results may be considered as justification for combining BIB and pace estimates of the distribution of a generalized reading proficiency variable. Defined operationally at each age level, this variable implies performance on the NAEP items under BIB administration conditions through the three-parameter logistic IRT model and the BIB item parameters, and under pace conditions through the same model after the linear transformation that matches the first two moments of the appropriate age sample.

After the translation has been accomplished, any characteristic of the Year 15 age populations can be estimated by combining the nearly independent estimates calculated from the BIB and pace data. Consider, for example, two estimates $\overline{X}_B$ and $\overline{X}_P$ of the same subgroup mean calculated from

BIB and pace data, with corresponding jackknife variance estimates V_B and V_P. The combined estimate of the subgroup mean is

$$\bar{X}. = W_B\bar{X}_B + W_P\bar{X}_P$$

where

$$W_B = \frac{V_B^{-1}}{V_B^{-1} + V_P^{-1}} \quad \text{and} \quad W_P = \frac{V_P^{-1}}{V_B^{-1} + V_P^{-1}} \quad ;$$

its estimated variance is

$$V. = W_B^2V_B + W_P^2V_P \quad .$$

Figure 10.3-21

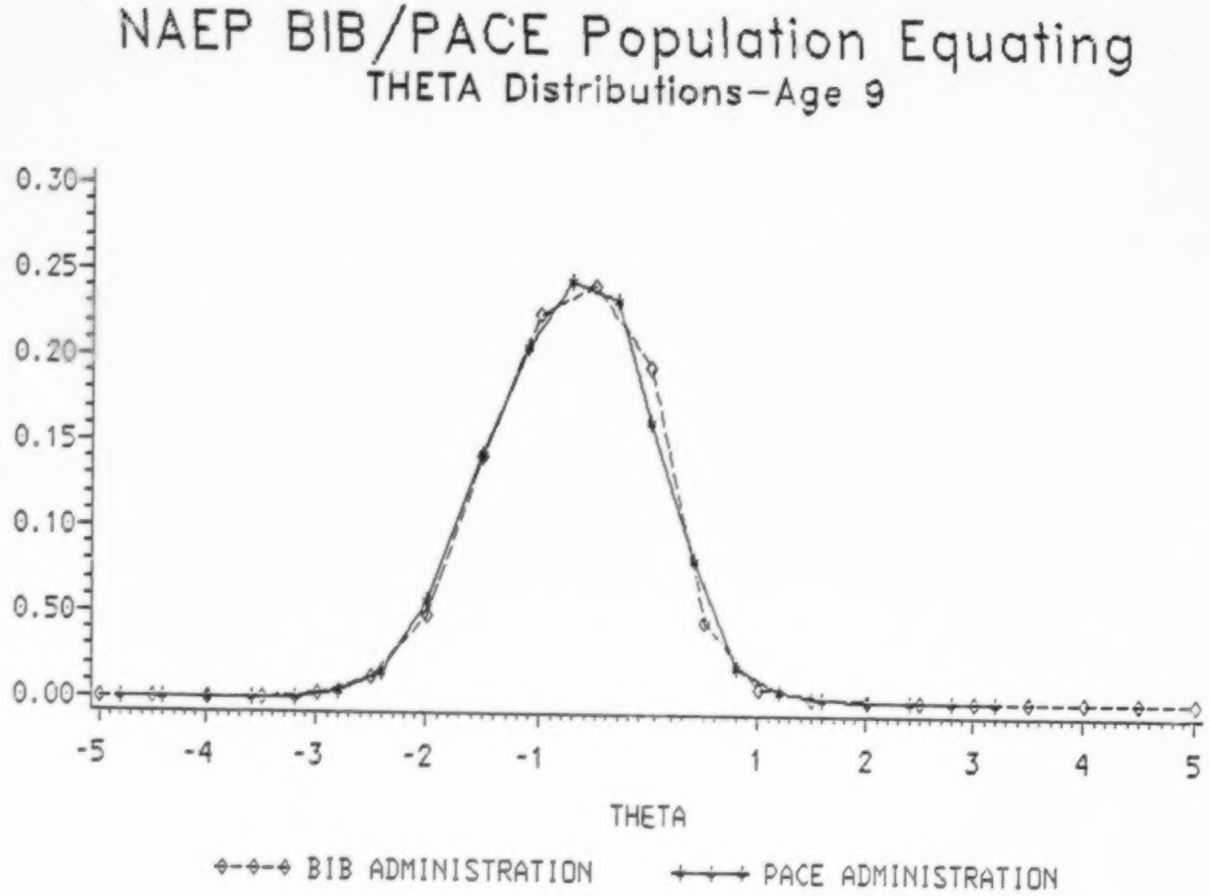


Figure 10.3-22

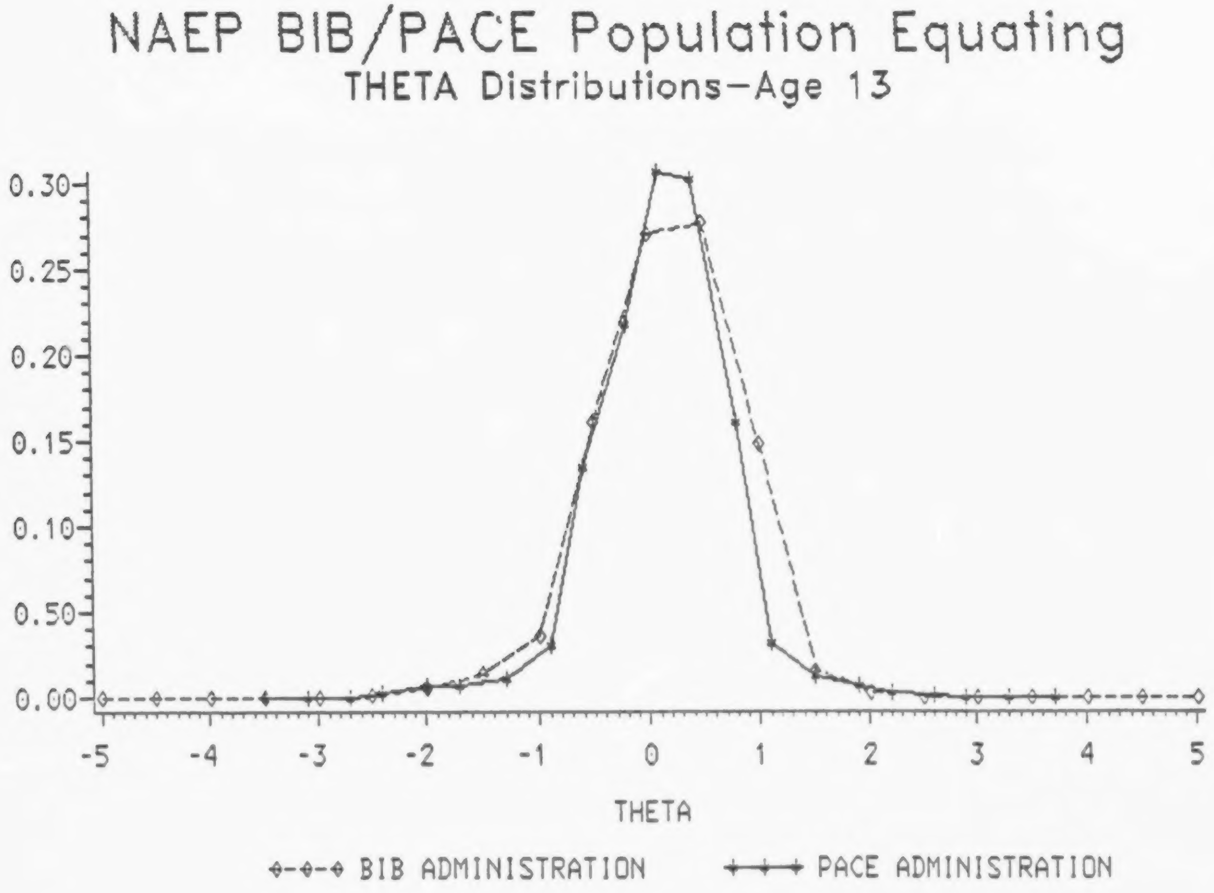


Figure 10.3-23

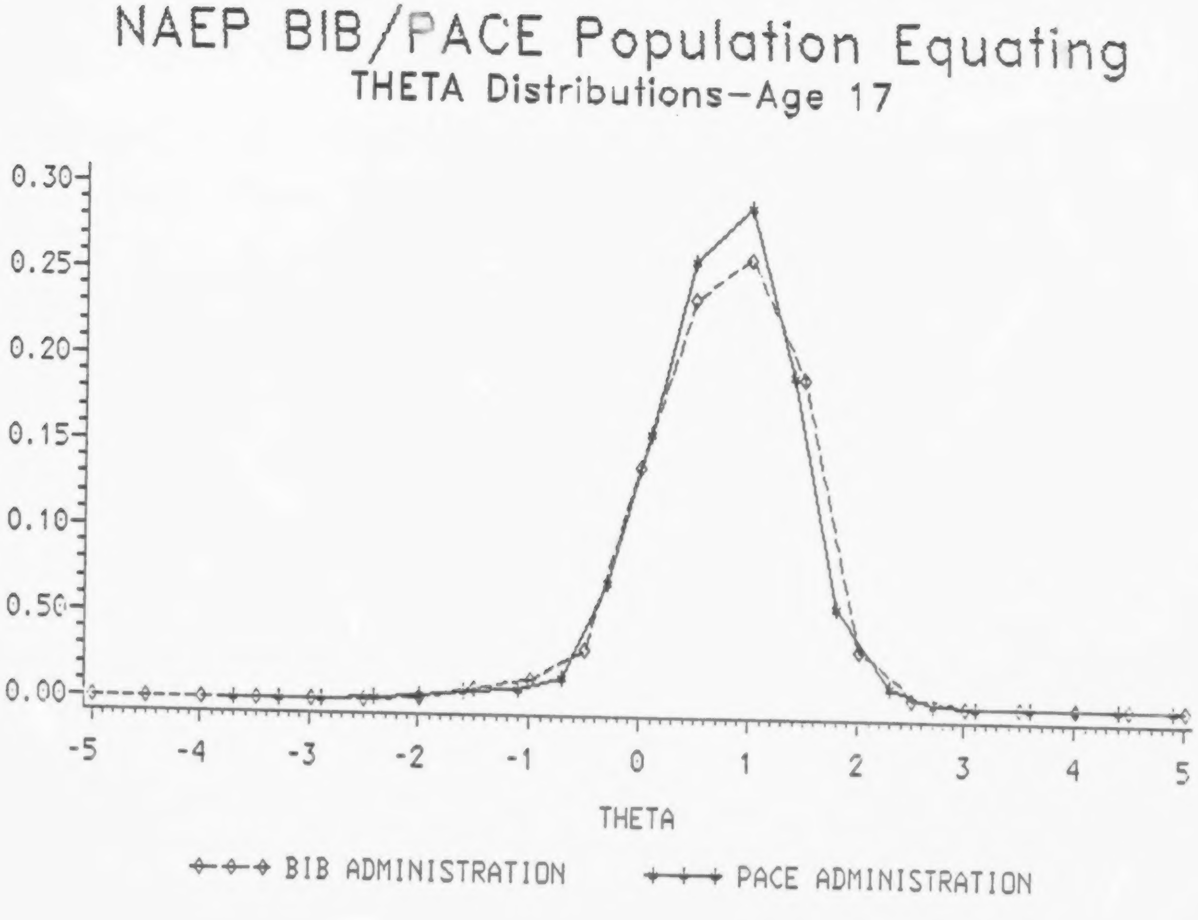


Table 10.3(12)
BIB/Pace Subgroup Means - Age 9

SUBGROUP		BIB	PACE
TOTAL		213.3	213.5
GENDER	Male	210.3	210.2
	Female	216.2	216.7
ETHNICITY	White	220.1	220.6
	Black	189.0	186.9
	Hispanic	193.6	191.2
	Other	222.0	227.0
REGION	NE	217.3	217.6
	SE	207.9	205.1
	Central	217.4	217.8
	West	210.9	213.6
PARENTAL EDUCATION	<H.S.	195.6	204.4
	Grad HS	211.3	212.4
	Post HS	224.4	224.2
	IDK	207.1	205.6
	Missing	177.5	174.0
STOC	Rural	205.1	205.5
	Low Met	194.2	197.1
	High Met	232.3	230.7
	Big City	210.6	212.8
	Fringe	214.4	215.2
	Med City	213.4	212.5
	Small Place	214.5	214.8
ABOVE, AT, OR BELOW MODAL GRADE LEVEL	< Modal Grade	187.4	187.6
	= Modal Grade	221.7	225.5
	> Modal Grade	250.5	254.6
	Missing	0.0	0.0
ITEMS IN HOME	< 3 Items	201.3	200.9
	= 3 Items	217.3	217.7
	= 4 Items	224.8	227.7
	IDK	177.8	155.6
	Missing	179.2	174.9
TV	0-2 Hours	220.4	218.6
	3-5 Hours	219.5	221.7
	6-More	201.6	203.9
	Missing	205.2	192.5

Table 10.3(13)

BIB/Pace Subgroup Means - Age 13

SUBGROUP		BIB	PACE
TOTAL		258.3	258.0
GENDER	Male	253.9	253.1
	Female	262.7	263.1
ETHNICITY	White	263.7	263.8
	Black	237.3	236.7
	Hispanic	241.4	237.4
	Other	262.2	262.6
REGION	NE	261.2	261.4
	SE	257.4	256.4
	Central	258.8	260.0
	West	256.1	254.7
PARENTAL EDUCATION	< H.S.	241.7	241.2
	Grad HS	253.3	255.3
	Post HS	269.6	267.2
	IDK	237.3	237.4
	Missing	255.0	254.7
STOC	Rural	255.8	256.0
	Low Met	239.5	239.4
	High Met	275.7	273.4
	Big City	255.2	254.8
	Fringe	260.7	260.0
	Med City	257.6	257.8
	Small Place	258.3	258.4
BELOW, AT OR ABOVE MODAL GRADE	< Modal Grade	240.0	238.7
	= Modal Grade	266.5	266.9
	> Modal Grade	300.6	295.8
	Missing	0.0	0.0

Table 10.3(13)
(continued)

SUBGROUP		BIB	PACE
ITEMS IN HOME	< 3 Items	240.6	243.6
	= 3 Items	256.1	255.1
	= 4 Items	266.2	265.1
	IDK	248.1	194.2
	Missing	250.4	249.2
TV	0-2 Hours	267.1	267.5
	3-5 Hours	262.2	262.4
	6-More	245.8	247.3
	Missing	243.7	240.2
HOMEWORK	Had None	255.5	256.6
	Didn't Do	246.5	248.8
	< 1 Hour	261.1	261.1
	1-2 Hours	265.6	266.4
	> 2 Hours	263.7	262.8
	Missing	238.5	232.9

Table 10.3(14)

BIB/Pace Subgroup Means - Age 17

SUBGROUP		BIB	PACE
TOTAL		288.1	288.3
GENDER	Male	282.7	284.6
	Female	293.7	292.1
ETHNICITY	White	293.9	295.4
	Black	265.1	259.8
	Hispanic	269.1	268.2
	Other	288.4	283.0
REGION	NE	289.8	291.9
	SE	285.3	282.1
	Central	289.2	289.3
	West	287.7	290.0
PARENTAL EDUCATION	< H.S.	269.8	268.7
	Grad HS	280.5	280.7
	Post HS	299.7	300.6
	IDK	256.8	255.2
	Missing	292.1	265.3
STOC	Rural	283.0	282.0
	Low Met	265.8	266.2
	High Met	300.4	301.3
	Big City	288.2	287.2
	Fringe	289.8	294.4
	Med City	291.0	288.2
	Small Place	287.7	288.2
BELOW, AT, OR ABOVE MODAL GRADE FOR AGE	< Modal Grade	260.3	256.9
	= Modal Grade	294.8	295.0
	> Modal Grade	304.3	300.7
	Missing	0.0	0.0

Table 10.3(14)
(continued)

SUBGROUP		BIB	PACE
ITEMS IN HOME	< 3 Items	266.5	267.1
	= 3 Items	282.7	284.7
	= 4 Items	294.8	294.4
	IDK	233.5	229.4
	Missing	274.8	224.9
TV	0-2 Hours	295.0	295.6
	3-5 Hours	283.7	286.1
	6-More	269.3	273.3
	Missing	255.9	253.9
HOMEWORK	Had None	276.8	280.4
	Didn't Do	286.9	290.2
	< 1 Hour	288.9	289.5
	1-2 Hours	293.5	293.0
	> 2 Hours	299.6	297.1
	Missing	245.9	241.7

Figure 10.3-24

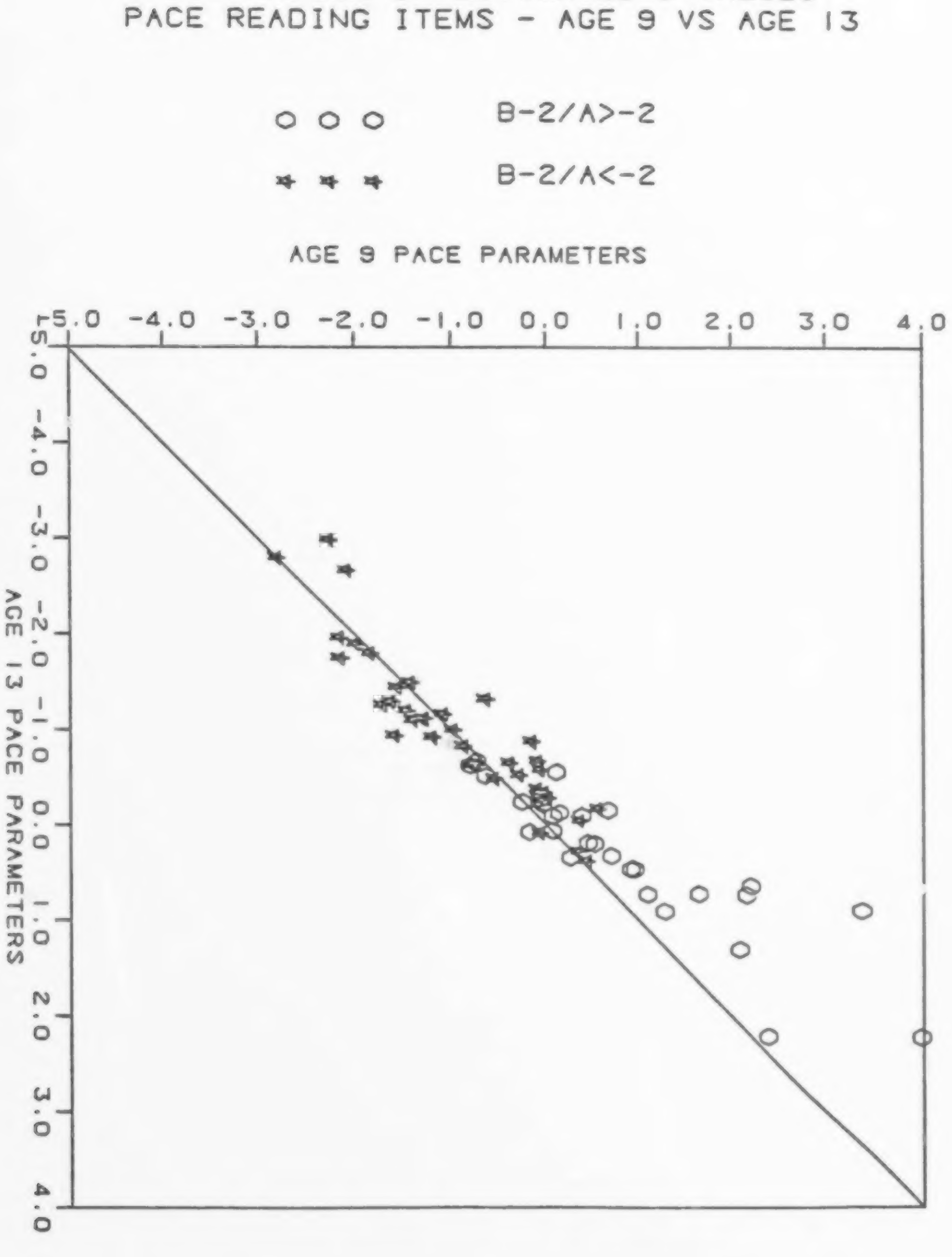
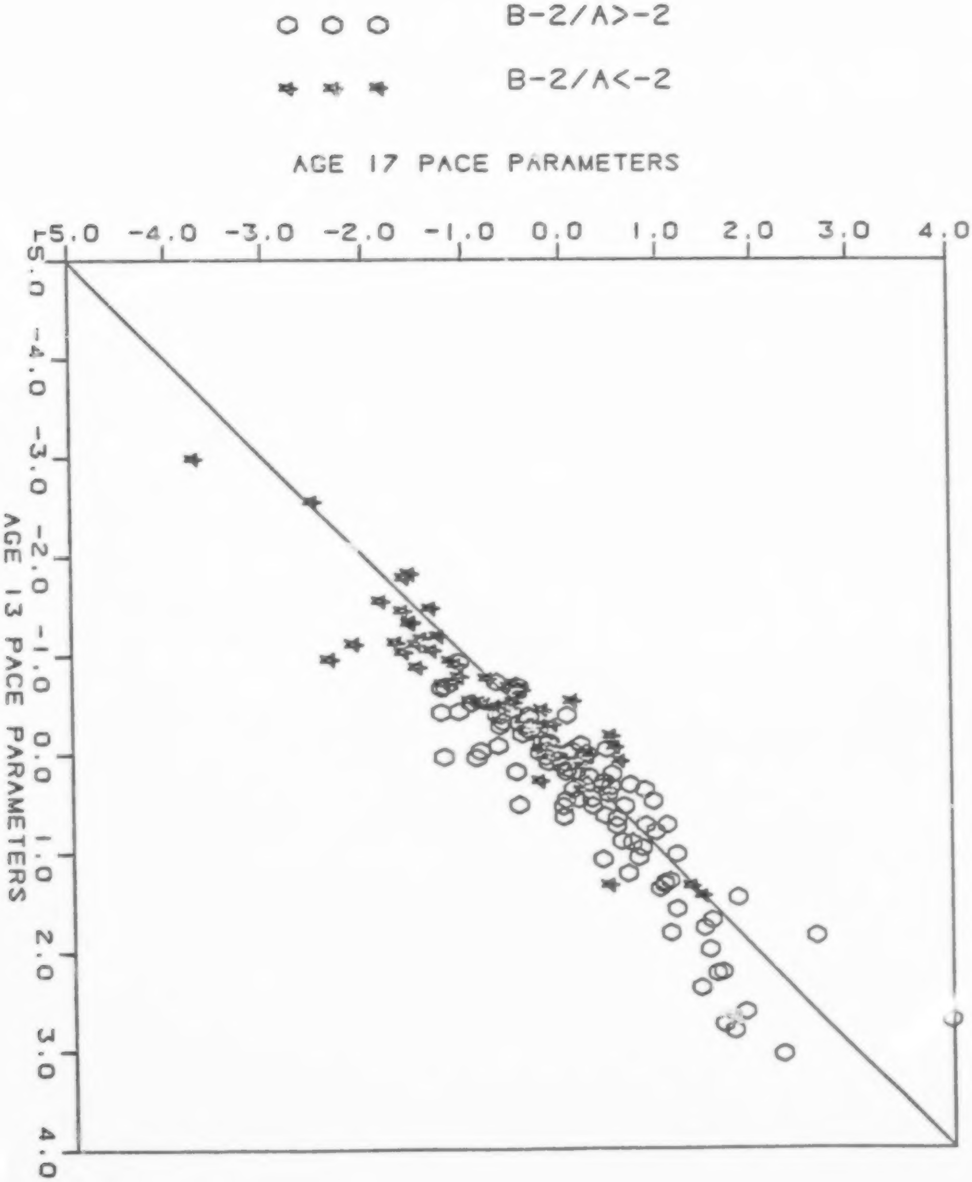


Figure 10.3-25

COMPARISON OF ESTIMATED B VALUES
PACE READING ITEMS - AGE 17 VS AGE 13



Chapter 10.4

TREND ANALYSIS

Robert J. Mislevy
Kathleen M. Sheehan

Educational Testing Service

Tracking trends of reading proficiency over time, as well as offering comparisons among subpopulations at a given time, are major objectives of the National Assessment. In this section, we summarize procedures taken toward these ends with NAEP reading data. Attention is focused on IRT procedures. We describe the manner in which data from past reading assessments was selected for analysis, item calibration procedures (again under the three-parameter logistic model), the estimation of effects for historically important NAEP reporting variables, and the generation of plausible values. These procedures were carried out on responses solicited under paced tape administration, including the Year 15 paced tape bridging study. The methods used to place these results on the Year 15 BIB reading proficiency scale were described in Section 10.3.6.

10.4.1 Estimation of Trends

To obtain maximum information about trends in reading proficiency over time, separate trend lines were estimated for each age group. Each within-age analysis was conducted using data from the 1983-84 assessment and from three previous assessments. The previous assessments were conducted by the Education Commission of the States (ECS) during the 1970-71, 1974-75, and 1979-80 school years. These three assessments are referred to as Years 2, 6 and 11 in the technical discussions which follow. The 1983-84 assessment, conducted by Educational Testing Service, is referred to as Year 15. All assessments prior to Year 15 were conducted under paced tape administration conditions. The Year 15 assessment was conducted under both paced tape and BIB spiral conditions. To avoid confounding pacing effects with time effects, the paced tape and BIB spiral data collected in Year 15 were analyzed separately.

The original analysis plan was to estimate item parameters for BIB data and trend data separately using LOGIST, to link the scales on the basis of LOGIST item parameter estimates, then obtain maximum likelihood estimates (MLEs) for the ability of each sampled pupil with these item parameter estimates via the MLE-ABIL program (M.S. Wingersky, 1984).

As noted in the preceding chapters, LOGIST succeeded in providing item parameter estimates from the BIB data. MLEs were superseded by plausible values, however, due to problems with infinite MLEs. At this point, plausible values computed from LOGIST item parameters characterized the BIB data.

As a precursor to estimating trend item parameters with LOGIST, an analysis of the number of items linking assessment booklets in previous reading assessments was carried out. Such links are LOGIST's basis of determining a common scale across test forms. The analysis revealed very weak links among the booklets of a given age in a given year; often only three or four linking items were present, and their content was not necessarily representative of the assessment as a whole. Another source of information for linking that was available, however, lies in the fact that with appropriate case weights, the samples administered each booklet in a given age/year were randomly equivalent samples from the same population. BILOG is able to incorporate this linking information, and was therefore selected for item calibration in the trend data. For reasons discussed in Section 10.3.6, separate analyses were carried out for each age.

Plausible values were then computed for trend data in essentially the same manner as described for BIB data in the preceding section. By the process described in Section 10.3.6, an equating of the 1984 BIB age samples (with LOGIST item parameters) and pace samples (with BILOG item parameters) was carried out. The checks described in Section 10.3.6 proved unsatisfactory. It was hypothesized that the differences were due to the use of LOGIST item parameters for one sample and BILOG for the other. After BIB item parameters were re-estimated with BILOG (Section 10.3.2), the linking of BIB and pace did prove more satisfactory. Plausible values based on BILOG item parameters thus provide the data on trends that were reported in The Reading Report Card: Progress Toward Excellence in Our Schools (1985).

10.4.2 Selection of Items and Forms to Include

To maximize year-to-year linkages while minimizing the total number of items calibrated, only those booklets which contained relatively high proportions of items in common with other assessment years were included in the trend analysis. Table 10.4(1) lists the booklets selected for each grade/age. For Year 2, the eight booklets which provided the most linking items were selected. Since only three booklets were administered in Year 6, all three were selected. For Year 11, only those booklets which contained five or more items in common with the Year 15 data were selected. All of the booklets administered in pace format in Year 15 were selected. These booklets provided a total of 633 distinct verbal items: 496 reading items and 137 study skills items. The reading items were used in the IRT-based trend analysis which is documented here. Throughout this chapter, items are identified by a three digit number which indicates their position on the trend data file. Table B-1 in Appendix B provides additional identification information for each item, including both the ECS ID number and the ETS ID number, where appropriate.

Items which had been administered by both ECS and ETS were examined to ensure equivalence across administrations. This investigation revealed 33 reading items which had undergone significant changes in format between the ECS and the ETS administrations. These 33 items were treated as new items in Year 15. Table 10.4(2) lists these 33 "changed" items along with their old and new item numbers.

The addition of 33 "new" reading items to the trend data yielded a total of 529 effective items for the IRT trend analysis. Of this total, 217 were administered to 9-year-olds, 211 were administered to 13-year-olds, and 205 were administered to 17-year-olds. These data were screened for anomalous patterns in proportion-correct across administration years. Table 10.4(3) lists proportion-correct data for eleven items which were flagged by the screening procedure. Several of these items were eventually excluded from the analysis. Excluded items are noted in the table by an N (Not Calibrated). Items retained in the analysis are noted by a C (Calibrated). Supporting evidence for the decision to include or exclude each "questionable" item follows.

Three of the items listed in Table 10.4(3) (#119, #145, and #368) were flagged because they showed a sudden drop in proportion-correct in Year 15. Further investigation of these items revealed that all three contained significant format changes which had been previously overlooked. These three items were excluded as noted. Five items which were not administered in Year 15 showed unexplained shifts in the proportion of correct responses in Years 2, 6, and 11. These items (#3, #142, #150, #153, and #175) were also excluded. (It is useful to note that Item #3 was the only item excluded from the analysis of the BIB data. In the BIB data file, this item was coded as Item #20.) Item #105 was excluded from the Age 17 analysis because it was too easy to provide any information about the population distribution; in the calibration sample of at least 1,000 subjects responding to the item, its proportion-correct value was 1.0. Item #251 was excluded from the Age 9 analysis for the same reason. Item #179 was flagged because the calculated proportion-correct for Age 9 was higher than that for Age 13. This item was found to have been mis-keyed and was excluded from all trend analyses. In total, five items were excluded from the Age 9 trend data, seven items were excluded from the Age 13 trend data and three items were excluded from the Age 17 trend data.

10.4.3 Estimation of Trend Item Parameters

The BILOG computer program was used to estimate item parameters for the trend data. Separate calibrations were performed for each age group. A random sample of examinees was selected for each calibration. The sampling frame consisted of all examinees who were administered the booklets listed in Table 10.4(1). The sample selected for each age group provided approximately 1,000 examinees for each item for each year. The modified BILOG priors described in Section 10.3.2 were found to be appropriate for all three age groups. Each "Age-Only" calibration required approximately twelve EM cycles and one Newton step.

Table 10.4(1)

Booklets Selected for Calibrating Trend Items

Age	Year	Booklets
9	2	1,2,3,4,5,6,7,9
	6	1,2,3
	11	1,2,3,4,10,11
	15	1,2,3,4

Age	Year	Booklets
13	2	1,2,3,4,5,11,12,13
	6	1,2,3
	11	1,2,3,14
	15	1,2,3,4

Age	Year	Booklets
17	2	2,3,4,5,7,8,9,10
	6	1,2,3
	11	1,2,3,11
	15	1,2,3,4

Table 10.4(2)

"Changed" Reading Items

New Item No.	ECS ID	ETS ID	Old Item No.
634	H284000-A002/002	N003202	445
635	H284000-A003/003	N003203	446
636	7102010-A001/001	N005101	92
637	7099007-A001/001	N005001	12
638	7103044-A002/002	N001702	142
639	7103044-A003/003	N001703	143
640	H265000-A003/003	N002003	435
641	7503045-A001/001	N003301	386
642	7303019-A002/002	N001202	289
643	7401016-A001/001	N004801	342
644	7103020-A001/001	N003901	117
645	H413000-A001/001	N008201	494
646	H413000-A005/005	N008205	498
647	H222000-A004/004	N001603	407
648	7127001-A002/002	N003802	170
649	7127001-A003/003	N003803	171
650	7099006-A001/001	N004201	10
651	7099006-A002/002	N004202	11
652	7127003-A002/002	N002102	175
653	H442000-A001/001	N008108	519
654	H241000-A002/002	N004402	418
655	H241000-A003/003	N004403	419
656	H404000-A004/004	N013104	472
657	H416000-A002/002	N010002	499
658	7503001-A001/001	N011101	379
659	H265000-A001/001	N010501	398
660	7102008-A001/001	N010301	91
661	H201000-A002/002	N008602	388
662	H405000-A001/001	N001501	473
663	H405000-A002/002	N001502	474
664	H405000-A004/004	N001504	476
665	7401071-A001/001	N010201	355
666	H287000-A001/001	N013501	451

Table 10.4(3)

Proportion Correct for Questionable Items

Item[1]	Age[2]	Year 2	Year 6	Year 11	Year 15	Status[3]
3	9	0.1863	0.2095	0.1642		N
	13	0.1801	0.1273	0.1233		N
105	17	1.0000				N
119	9	0.8646	0.8094	0.8286	0.5863	N
	13	0.9353	0.9068	0.9165		C
142	13	0.4305	0.3188	0.3408		N
	17	0.4799	0.4611	0.4533		C
145	13		0.5847	0.5904	0.4794	N
	17	0.7225	0.8142	0.7444	0.6096	N
150	9	0.6962	0.6594	0.7827		C
	13	0.8617	0.9311	0.9210		N
153	13	0.1531	0.0695	0.0831		N
	17	0.2349	0.1580	0.1329		C
175	9	0.1177	0.1538	0.1158		C
	13	0.5559	0.4383	0.4416		N
	17	0.5812	0.7061	0.5724		N
179	9	0.1288				N
	13	0.0451				N
251	9	1.0000				N
368	9	0.6153	0.5972	0.7135	0.4990	N

[1] The item number is the position of the item on the trend data file.

[2] Only age groups to which the item was administered are listed.

[3] Items marked N were administered but not calibrated; items marked C were administered and calibrated.

Diagnostic plots were produced after every fourth cycle. A sample diagnostic plot is given in Figure 10.4-1. As in the figures presented earlier, the smooth line is the fitted three-parameter logistic item response curve and the points represent expected proportions of correct responses for various subgroups of examinees. In these particular plots, examinees are classified according to the calendar year in which they were tested.

These plots revealed six poorly fitting items which were later excluded from the trend analysis. Table 10.4(4) identifies each item and indicates which particular "Age-Only" analysis was affected. Diagnostic plots for these six items are given in Figure 10.4-2.

The number of items included in the final trend calibrations is provided in Table 10.4(5). The final item parameter estimates and corresponding standard errors are given in Tables B-2 through B-4 in Appendix B. These parameter estimates were originally estimated on a provisional scale but were re-scaled so that, for each age group, the distribution of reading ability in the Year 15 pace sample would have the same first two moments as the distribution of reading ability in the BIB sample. Ability distributions were estimated using RESOLVE (Mislevy, 1985c). The trend item parameters were re-scaled in this manner so that the results of the trend analysis could be reported on the Reading Proficiency Scale. (Additional details on the Year 15 BIB/pace linkage are provided in Section 10.3.5).

Tables B-5 through B-7 in Appendix B provide item linkage information for all of the items included in the final BILOG calibrations. This information includes:

(1) the total number of items calibrated in each booklet;
(2) the number of calibrated items linking booklets across years; and
(3) the number of calibrated items linking booklets within years.

The position of each item within its test booklet is also provided. (These numbers appear in the columns of the tables.)

10.4.4 Estimation of Conditional Effects

Conditional distributions of reading proficiency given background responses were estimated separately for each age group and for each assessment year. Background variables were chosen to be as similar as possible to the background variables used in the analysis of the Year 15 BIB data.

One change that could not be avoided involved the definition of examinee ethnicity. In the Year 15 assessment, information about examinee ethnicity was available from a variety of different sources. This information was combined to form a derived variable, labeled "imputed race/ethnicity," which was used in the analysis of the Year 15 BIB data.

Figure 10.4-1

Diagnostic Plot for Item 87
(Calibrated for Examinees in Grade 4/Age 9)

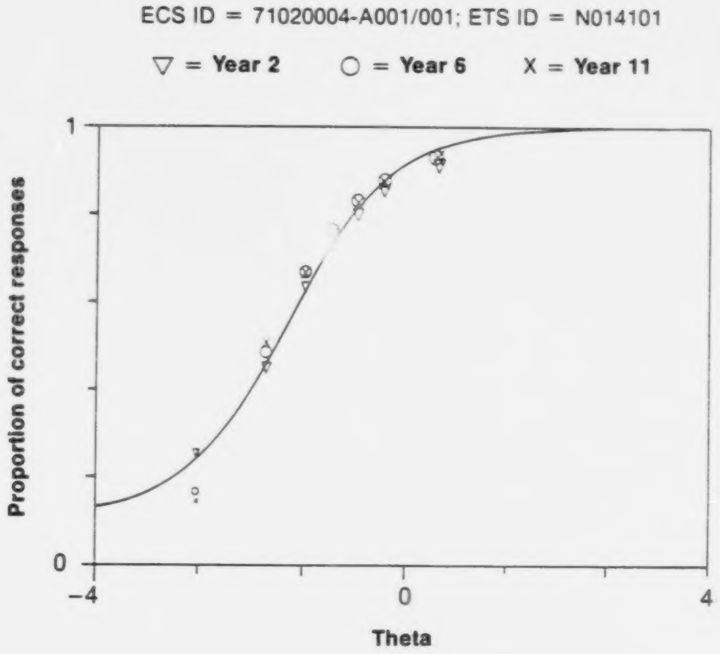


Figure 10.4-2

Plots of Items Excluded During Preliminary Calibrations of Trend Data

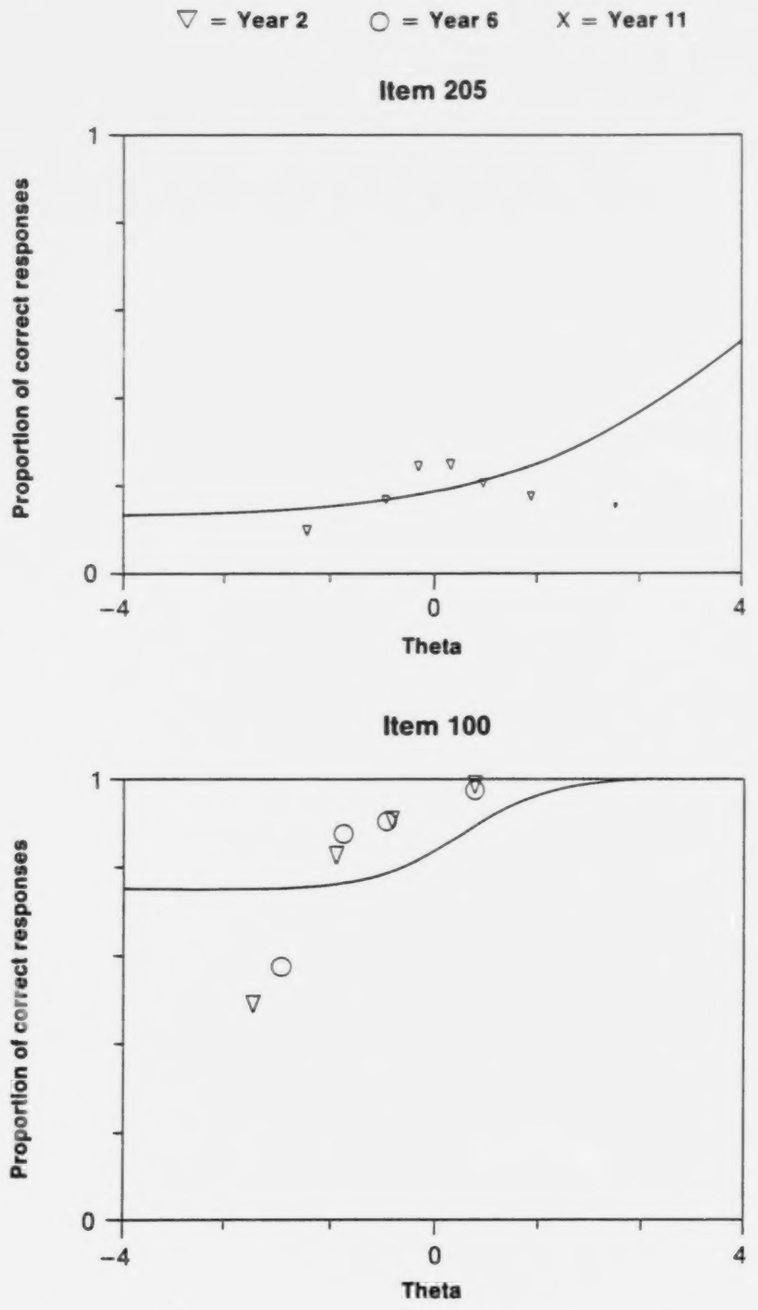


Figure 10.4-2
(continued)

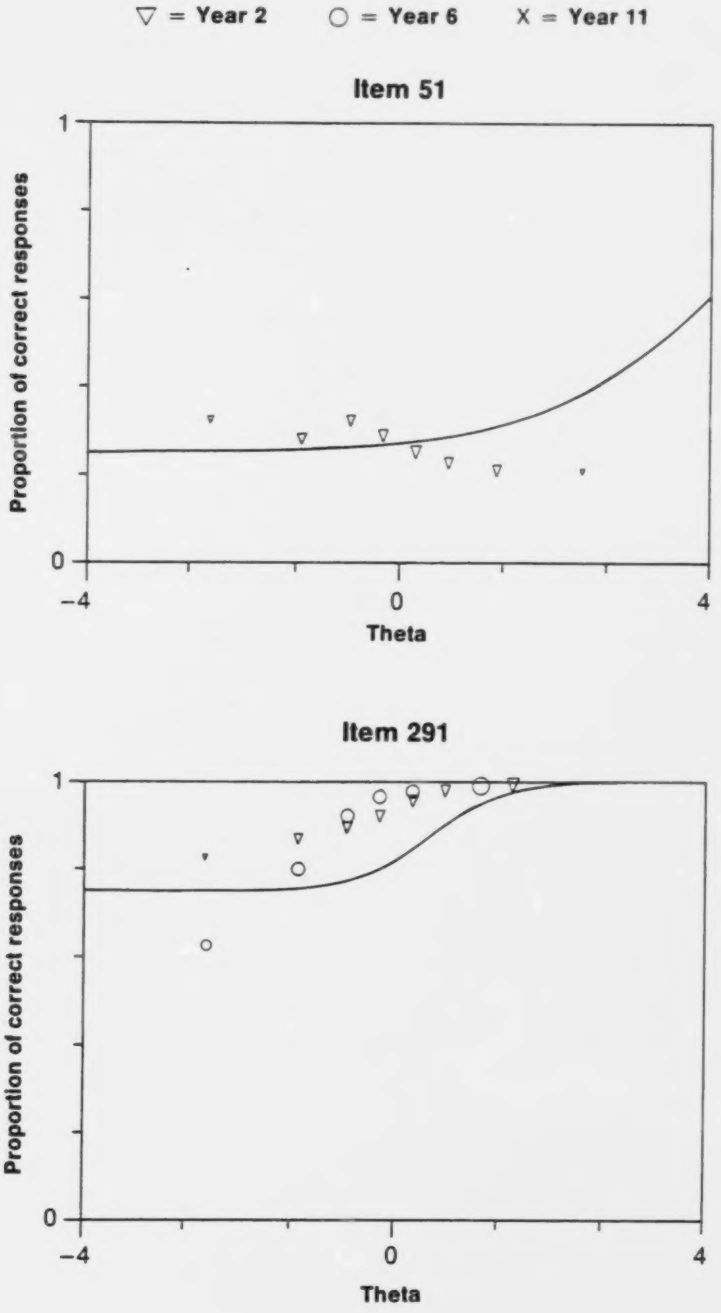


Figure 10.4-2
(continued)

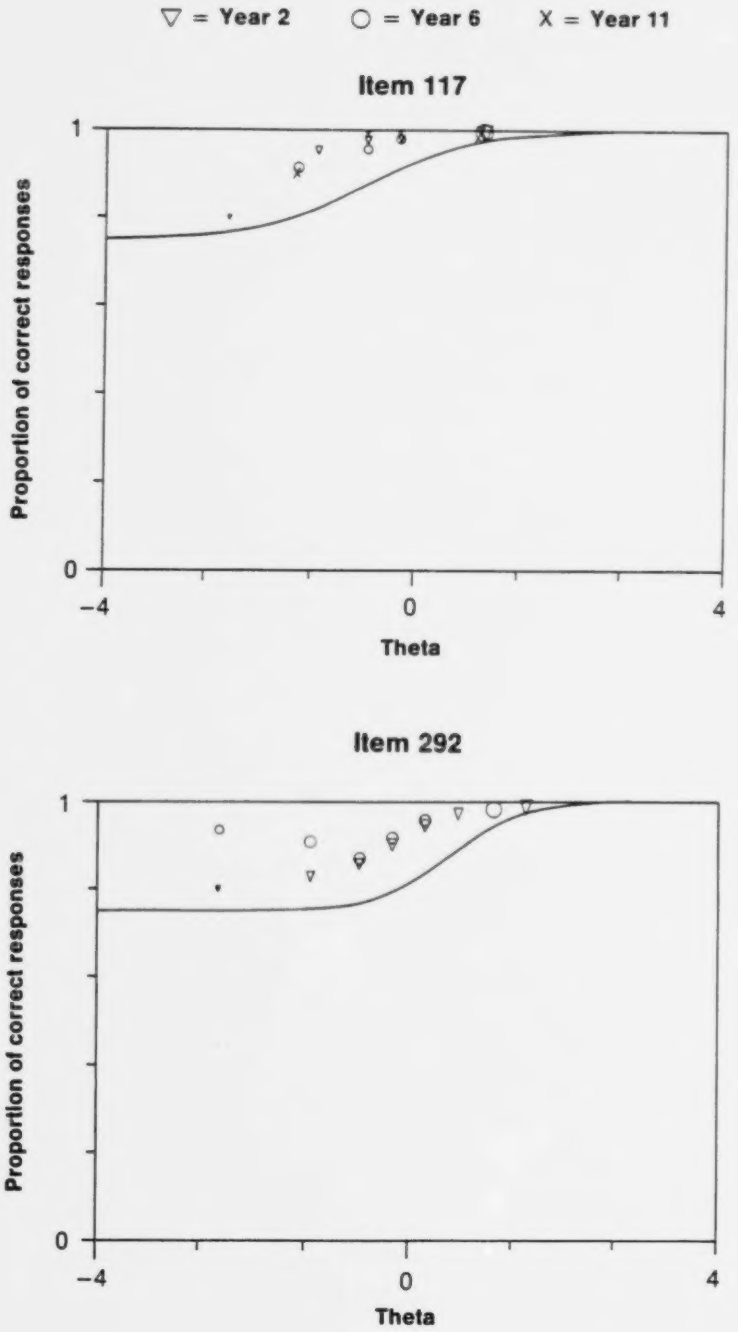


Table 10.4(4)

Items Excluded During Preliminary Calibration Runs*

Item	Grade 4/ Age 9	Grade 8/ Age 13	Grade 11/ Age 17
51		C	N
100	N		
117		N	
205	C	N	C
291		C	N
292		C	N

*C = administered and calibrated; N = administered but not calibrated.

Table 10.4(5)

Item Calibration Summary

	Number of Potential Trend Items	Number Excluded During Screening	Number Excluded After Initial Calibration	Number Included In Final Calibration
Grade 4/Age 9	217	5	1	211
Grade 8/Age 13	211	7	2	202
Grade 11/Age 17	205	3	3	199

(The exact definition of the "imputed race/ethnicity" variable can be found in Section 12.1.) However, because the only type of ethnicity information available from previous NAEP assessments was observed ethnicity, the conditioning variable used for the trend analysis was "observed ethnicity" rather than "imputed race/ethnicity."

The variable coding scheme developed for the data collected in Years 6 and 11, and the Year 15 pace data, mirrored the scheme developed for the Year 15 BIB data with one exception: the grade/age variable was coded with three levels rather than five. These three levels were defined as:

Level	Description
1	at Modal age, < Modal grade
2	at Modal age, = Modal grade
3	at Modal age, > Modal grade

This same three-level grade/age variable was included in the coding scheme developed for the Year 2 data. Two additional changes were also incorporated into the coding scheme developed for the Year 2 data:

(1) The ethnicity effect was coded with two levels rather than three, because in Year 2, Hispanics were not coded as a separate ethnic group. The two coded levels for ethnicity were

Level	Description
1	Black
2	White and Other (Including missing)

(2) The Region effect was excluded, because the Region variable was incorrectly coded in Year 2.

The dataset used to estimate conditional effects included all examinees who were administered a booklet containing at least two calibrated items. This dataset included some examinees who were administered booklets which were not used for item calibration but which did include two or more items which also appeared in booklets which were used for item calibration. These additional booklets are listed in Table 10.4(6). Examinees who did not respond to any of the calibrated items were included on the file but were not used by the estimation procedure. The estimated effects, and sample sizes, are given in Tables 10.4(7) through 10.4(10).

10.4.5 Generation of Plausible Values

Five plausible values were generated for each examinee who was administered at least one of the booklets listed in Tables 10.4(1) and 10.4(6). The methodology used to generate the plausible values exactly parallels the methodology which was used to generate plausible values for the Year 15 BIB data. This methodology is described in Section 10.3.4. The file of plausible values produced was used to estimate the trend lines which were reported in The Reading Report Card (1985).

Table 10.4(6)

Additional Booklets* Used For Estimating Conditional Distributions

Age	Year	Booklets
9	2	8
	11	6,8

Age	Year	Booklets
13	2	6,8
	11	4,6,13,15

Age	Year	Booklets
17	2	1,6
	11	4,13,14

*These booklets were not used to estimate item parameters. However, each contained two or more items which also appeared in a booklet which was used to estimate item parameters.

Table 10.4(7)

Estimated Conditional Effects
Year 2 Pace Data

Effect	Level	Age 9	Age 13	Age 17
Intercept	All subjects	-2.570618	-1.182882	-1.087587
Sex	Female	0.205023	0.201680	0.171321
Ethnicity	White and Other	0.708416	0.638349	0.696148
STOC	High Metro	0.460787	0.088009	0.314084
	Not High or Lo Metro	0.201538	-0.071709	0.127357
Parental Ed.	High School	0.282657	0.232633	0.227909
	Beyond HS	0.489658	0.437952	0.484824
	All else	0.108501	-0.030201	0.065427
Grade/Age	= M age, = M grade	0.702134	0.581215	0.740842
	= M age, > M grade	1.137812	0.904846	0.911997
Misc.	Subjects with unrecoverable missing values.	0.448089	-0.257232	-1.898172

	Age 9	Age 13	Age 17
Number of Examinees	18,096	23,938	18,417
Estimated Variances	0.4631	0.30544	0.45796

Table 10.4(8)

Estimated Conditional Effects
Year 6 Pace Data

Effect	Level	Age 9	Age 13	Age 17
Intercept	All subjects	-2.334113	-1.295558	-0.905302
Sex	Female	0.195583	0.221243	0.154221
Ethnicity	White and Other	0.539376	0.525777	0.639462
	Hispanic	0.112853	0.281520	0.374550
STOC	High Metro	0.440385	0.355996	0.290733
	Not High or Lo Metro	0.278599	0.216792	0.173098
Region	Central	-0.135823	-0.041520	-0.105375
	South East	0.067479	0.083684	0.008417
	West	-0.069881	-0.043541	-0.106991
Parental Ed.	High School	0.273121	0.189764	0.166982
	Beyond HS	0.413783	0.427720	0.433708
	All else	0.124897	-0.023117	-0.201775
Grade/Age	= M age, = M grade	0.625358	0.512917	0.649586
	= M age, > M grade	1.035986	0.833832	0.810016
Misc.	Subjects with unrecoverable missing values.	1.223144	-0.159365	0.858545

	Age 9	Age 13	Age 17
Number of Examinees	21,697	21,393	19,624
Estimated Variances	0.39646	0.32585	0.38421

Table 10.4(9)

Estimated Conditional Effects
Year 11 Pace Data

Effect	Level	Age 9	Age 13	Age 17
Intercept	All subjects	-2.114314	-1.032202	-0.867096
Sex	Female	0.159384	0.145401	0.114560
Ethnicity	White and Other	0.451578	0.491249	0.611792
	Hispanic	0.081409	0.216617	0.349791
STOC	High Metro	0.481785	0.252162	0.277156
	Not High or Lo Metro	0.270451	0.066009	0.168417
Region	Central	-0.098071	-0.057445	-0.046963
	South East	-0.005778	0.083751	-0.011844
	West	-0.088577	-0.067965	-0.042691
Parental Ed.	High School	0.228851	0.215626	0.140464
	Beyond HS	0.409781	0.490029	0.439863
	All else	0.100306	-0.076120	-0.102193
Grade/Age	= M age, = M grade	0.615816	0.433504	0.616226
	= M age, > M grade	1.187973	0.777249	0.792824
Misc.	Subjects with unrecoverable missing values.	0.740209	0.917640	0.246829

	Age 9	Age 13	Age 17
Number of Examinees	21,158	22,321	18,099
Estimated Variances	0.39041	0.33354	0.35543

Table 10.4(10)

Estimated Conditional Effects
Year 15 Pace Data

Effect	Level	Age 9	Age 13	Age 17
Intercept	All subjects	-1.850093	-0.908307	-0.537325
Sex	Female	0.098570	0.140171	0.119340
Ethnicity	White and Other	0.518862	0.425534	0.493804
	Hispanic	0.026952	0.076946	0.287789
STOC	High Metro	0.373506	0.279597	0.238584
	Not High or Lo Metro	0.150687	0.096120	0.174528
Region	Central	-0.143570	0.033821	-0.074830
	South East	0.000213	0.013760	-0.037671
	West	-0.037804	-0.021991	-0.016431
Parental Ed.	High School	0.040130	0.161891	0.123633
	Beyond HS	0.240621	0.328296	0.416555
	All else	-0.029917	0.013953	-0.146127
Grade/Age	= M age, = M grade	0.652164	0.473655	0.542676
	= M age, > M grade	1.328896	0.915722	0.665729
Misc.	Subjects with unrecoverable missing values.	1.159244	1.204793	1.443256

	Age 9	Age 13	Age 17
Number of Examinees	5,492	5,158	6,209
Estimated Variances	0.41479	0.33445	0.41769

Chapter 10.5

THE NAEP READING SCALE

Albert E. Beaton

Educational Testing Service

Since its inception, a major goal of the National Assessment of Educational Progress (NAEP) has been to report to decision makers at all levels what youths can and cannot do. To be useful, its reports should be psychometrically sound, yet easily interpretable; reports should be clear and concise, yet should not miss important subtleties of the learning area being assessed. The essential conflict between simplicity and detail requires careful thought, and decisions must be made about what information is most useful and what information can be judiciously excluded.

The NAEP staff has carefully considered how NAEP results would best be presented. The dimensionality of the Year 15 reading data was examined and it was found that much of the reading information could be summarized using a single dimension. Item response theoretic (IRT) methods were used as a way of estimating the item parameters for that reading dimension. Using the item parameters, sampling information, and the available information about individual students, estimates of the reading proficiency of American youth were made. After equating for differences in methods of administering items, reading data from the Year 2 (1970-71), Year 6 (1974-75), and Year 11 (1979-80) assessments were also scaled and population estimates were computed.

The purpose of this section is to describe the way that the NAEP reading results were presented. The next section will discuss the NAEP reading proficiency scale, which can be thought of as estimated true scores on a hypothetical test with known properties. After the scale is presented, we will discuss the anchoring of several scale points to specify what students at those points can and cannot do.

10.5.1 The NAEP Reading Scale

In its earlier years, NAEP reported educational progress by presenting the estimated percentage of students who responded correctly to each exercise. The percentages passing were also presented for selected subpopulations such as the different sexes, racial/ethnic groupings, and regions of the country. This approach allows a very detailed interpretation of what students can and cannot do. Insofar as the actual text of the exercises was made publicly available, a reviewer or policy maker could look at each exercise and, using its percent passing, judge the adequacy of student performance.

This approach soon proved to be cumbersome because too much detailed information was available for most policy makers to integrate and interpret. Some method of summarizing the information was clearly necessary. The past solution to the over-abundance of information was to publish the average of the percents correct over all exercises in a subject area such as reading. To avoid omitting too much detail, the average percent correct was also presented for sub-areas of interest; for example, in the reading assessment, the average percents correct were presented separately for literal comprehension, inferential comprehension, and reference skill exercises.

If all of the exercises had been administered to all of the students, then the average percent correct over all exercises would have been the same as the average percent correct over all students; that is, we could have reached the same value by computing the percent correct for each student and averaging over all students. The average percent correct could thus be considered as the average of the students' scores on a percent correct scale. However, it should be noted that the matrix sampling methods used in past and present NAEPs have the effect that all students do not receive the same exercises, so the average percent correct statistic is not precisely the same as averaging individual scores.

The average percent correct metric makes it awkward to report what students can and cannot do. First, the average percent correct metric depends on the selection of exercises; the selection of easy or difficult exercises could make student performance look good or bad, especially to a public that is accustomed to a "passing score" of 70 percent, for example. Second, since the metric is dependent on the selection of items, the items cannot be changed over time; items cannot be retired and replaced without changing the metric. This also restricts the ability to release exercises to the public. Third, age-to-age and grade-to-grade comparisons require that the same exercises be administered at all age or grade levels. Finally, even if all exercises were administered to all students, the average percent correct would not indicate what they could or could not do without examination of the individual exercise information.

Besides the percent correct metric, we considered and rejected a number of other reporting metrics. We did not want to present performance in a norm-referenced metric since the question was what students can and cannot do, not how they compare to each other or some norm group; thus, percentiles and grade equivalents were rejected. We did not want the metric easily confused with well known scales such as the SAT or ACT scales. And, of course, any scale that might be confused with IQ might mislead the casual reader about the assessment's meaning.

A seemingly simple way to proceed would have been to use the metric which is implicit in the IRT scaling procedures that were used. The LOGIST program produces a value called theta for each subject, and this value is an estimate of the subject's proficiency on the dimension being measured.

The BILOG procedure produces a distribution of plausible values in the same metric. Typically, the values of theta are standardized so that the average over all subjects is zero and the standard deviation is unity[1]. However, the theta scale results in negative scores which are more difficult for the public to interpret--and this might unduly affect a subgroup which received an average score below zero. Also, the theta scale is unbounded, with possible values anywhere from minus to plus infinity.

The LOGIST and BILOG programs can also produce an alternative score called the xi score. The xi score is the estimated true score on a test. For the Year 15 reading assessment, 228 scaled exercises were administered at one or more grades or ages. Thus, the xi scores would range from 48.7, chance level, to 228, the number of exercises. An advantage of the xi score is that it makes finite estimates possible for all subjects; those who respond to all exercises correctly are estimated to have perfect true scores on the test and those who did not do as well as chance are estimated at the chance level. Also, since the xi scores are like test scores, they are in a familiar type of metric.

However, using the xi scale would enshrine the particular reading assessment of Year 15 as the standard for all past and future reading assessments. These exercises were selected from a pool given by ECS, the previous grantee, to ETS, the present grantee. This set of exercises was not selected with any particular metric in mind; in fact, the set as a whole was relatively easy. The item parameters suggested unequal test information at different levels of the scale. Thus, reporting results as estimated scores on this particular assessment battery was rejected.

Instead of using the xi scale of the actual assessment battery, we chose to report the reading results as the estimated true score on a hypothetical reading proficiency test with somewhat idealized properties. The properties are as follows:

(1) The hypothetical test consists of 500 items. This property has the effect that test scores can range between zero and 500.

(2) All item characteristic curves are logistic, i.e. have the general form

$$p_{si} = 1/(1 + e^{-1.7(a(\theta_s - b_i))})$$

where p_{si} is the probability that a subject responds correctly to item i, a is the discrimination parameter, b_i is the difficulty parameter, and θ_s is the true proficiency score for subject s in the theta metric.

---

[1] LOGIST actually standardizes such that the standard deviation of the scores between -3 and +3 is unity.

(3) The correct answers to items cannot be achieved by guessing.

(4) All items discriminate equally; that is, $a = 1.5$ for all items. The value $a = 1.5$ was chosen since it is approximately the average value of the discrimination parameters for the actual items used in NAEP.

(5) Item difficulties are evenly distributed along the theta scale; that is, the b_i vary from -4.99 to +4.99 in steps of .02. Since almost all subjects will typically score in the range of -3.0 to +3.0, this condition means that the hypothetical test has about 100 items so easy that almost everyone responds correctly and about 100 items so difficult that they are failed by almost everyone.

Both Lord and Mislevy have shown that a scale defined in this way is essentially a linear function of the theta scale within the range of actual data. Holland and Zwick (1986) have provided a general function relating the theta scale to such hypothetical test scores. The particular function[2] used to translate from the theta scale to the reading proficiency (RP) scale was

$$RP_s = 250.5 + 50\,(\text{theta}_s)$$

where RP_s is the score of subject s on the reading proficiency scale.

Using this definition of a hypothetical test and since the distribution of theta has a zero mean and unit variance, we can make the following statements about the distribution of reading proficiency scores:

(1) The mean reading proficiency is 250.5 over all ages and grades combined.

(2) The standard deviation of proficiency scores is 50.

The NAEP RP scores ranged between about 75 and 425. The distribution of reading proficiency scores in NAEP is not normal, since three distinct ages and grades are included in the distribution. The overall distribution has three major modes, one for each grade/age combination, and there is considerable overlap among the distributions. Thus, the overall mean and standard deviation over all three grade/age combinations has little interpretive value.

---

[2] The function $RP_s = 250 + 50(\text{theta}_s)$ would have been preferable. Holland and Zwick (1986) have noted that the values actually used correspond to the b_i varying from -5.00 to +4.98 in steps of .02 instead of -4.99 to +4.99 as intended. The result is that the RP scores are a half-point higher than appropriate for the hypothetical test.

Clearly, the NAEP scale is not norm-referenced in the sense that knowing an individual's score by itself gives any useful information about how he or she compares to other individuals. The distribution of theta is used only to assure that the available exercises span the range, and a little bit more, where we expect students to be.

The hypothetical test would be appropriately fit by the Rasch model since there is no guessing and the item discrimination parameters are all identical.

The scale of the hypothetical test is equal interval; that is, if we constructed such a test then a subject who scored five points higher than another would be expected to answer five more items than the other no matter where on the scale the two subjects scored. In fact, the scale is, in a sense, a ratio scale since an estimated score of zero means that the subject would answer no items correctly; however, the zero is arbitrarily determined by the specified range of the difficulty parameters and a zero score does not mean that the subject has no reading proficiency at all.

Lord has noted that no one would or should build a test according to these specifications; having so many easy and difficult items would be inefficient. Also, a test for a particular purpose should have its items clustered near important decision points. However, we are using this hypothetical test for reporting purposes only and do not intend to attempt to construct a test with such properties.

10.5.2 Anchoring Scale Points

In this section we address the issue of presenting what students can and cannot do in reading. Our approach is to select a few points on the scale, find exercises that discriminate between what students at each point can do that students at lower levels cannot, and then attempt to generalize from the exercises to classes of competency.

If the reading items formed a perfect scale in the Guttman (1941) sense, then a person's test score would indicate exactly which items that person could answer correctly and which he or she could not. A score of 275, say, would indicate that the subject could answer the 275th item and all easier items but could not answer the 276th nor any more difficult item. If two subjects have distinct scores, then the two scores can be used to identify which items both can answer correctly, which items the higher scorer can answer and the other cannot, and which items neither can answer. At the item level, a Guttman scale immediately indicates what a person can and cannot do.

Of course, the NAEP reading exercises do not form a Guttman scale, and it is seldom that any real item response data have such ideal properties.

Multiple-choice items are especially unlikely to form a Guttman scale since the correct answer can be achieved by guessing. Also, subjects did not all receive the same items, which would complicate the interpretation of the Guttman scale.

For NAEP, we have searched the reading data for reading exercises which discriminate strongly between selected points on the scale, albeit these exercises are not perfectly discriminating as would true be for a Guttman scale. We have labeled these selected scale points by attempting to generalize from the highly discriminating exercises.

The general procedure used is as follows:

(1) Choose the scale points to anchor. The selection of the anchoring points is important since few items will be found that discriminate between close points and little useful information will be found if the points are far apart.

(2) Select items that discriminate between each pair of adjacent points. The following criteria were used for selecting items at each anchor point:

  (a) eighty percent or more of the students at that point could answer the item correctly.

  (b) less than 50 percent of the students at the next lower point could answer the item correctly. This criterion does not apply to the lowest valued anchor point.

  Using these criteria, an item can be selected for discriminating between only one pair of adjacent points.

(3) Batch the items found to discriminate between pairs of anchor points.

(4) Try to generalize from each batch of items to the level of accomplishment that the items represent. It is important that this step be performed by experts in the subject area.

(5) Try to understand the exercises that did not discriminate between any pair of points. Exercises may fail for a number of reasons such as measurement of another dimension, discrimination between points not chosen for anchoring, or, perhaps, simply poor item construction.

The details of the process as implemented for the NAEP are as follows:

(1) Anchor points were selected. For the NAEP reading scale, we chose to anchor the following points: 150, 200, 250, 300, and

350. These points span the range in which most subjects scored.

(2) The probability of obtaining correct responses to each of the NAEP reading proficiency exercises at each point was estimated. This step was done using the parameters of the item characteristic curves which were available from the LOGIST (Wingersky, Barton, & Lord, 1982) and BILOG (Mislevy & Bock, 1982) programs.

(3) The RP point at which the probability of passing was .80 was computed. This was computed using the item parameters and solving the equation of the three parameter logistic model for the value of theta for which $p_{*i} = .80$. The theta value for each item was then transformed to the RP scale. These values are called $RP80_i$.

Steps 2 and 3 were computed by an IBM-PC program called Behanc (Beaton, 1986).

(4) The items were sorted by RP80. This was done to place the items in an order of difficulty. The actual texts of the items were cut and pasted onto sheets of paper which were placed in a binder in RP80 order.

(5) The item statistics, including whether or not they met the anchoring criterion, were pasted into the item text book underneath the item text. Items meeting the discrimination criteria were highlighted.

(6) Red markers were placed in the item text book to pinpoint the item with RP80 value closest to each anchor point and blue markers were entered to mark the mid-points between anchor points (e.g., 175, 225, etc.),

(7) The item text books were delivered to reading consultants who were asked to interpret the items. The meaning of the item statistics was described. The panel of subject matter specialists were asked to look at the batches of items and describe what students at each level could do that students at lower levels could not. We asked for a description in a paragraph or two, then a summary sentence, and, finally, a one word label for the point. They were also asked to select several items, which met the criteria, to serve as exemplars for each anchor point.

(8) In developing the descriptions, the reading consultants used their expert judgment as well as descriptive statistics of the passage and item types to characterize the relationship between the type of question asked and the text characteristics.

(9) The anchor point descriptions, along with the item examples, were then reviewed by 25 additional reading specialists. The descriptions, sentences, and labels were revised incorporating suggestions of these reading specialists.

(10) Not all items failing to meet the discrimination criteria were studied to find out exactly why; these were given less priority in selecting items for the 1985-86 assessment.

The results of the anchoring process were published in The Reading Report Card: Progress Toward Excellence in Our Schools (1985, p. 15). The description of the anchor points is repeated here as Figure 10.5-1. The five levels of proficiency were defined as Rudimentary (150), Basic (200), Intermediate (250), Adept (300), and Advanced (350). The Reading Report Card also includes at least two sample items for each anchor point.

The anchoring process allows the description of what a student can and cannot do in terms of levels of reading performance, not in terms of what other students do or do not do. We can estimate directly the number or proportion of students who can perform at different levels, which is the sort of information needed for policy action. For example, such statements as 64.2 percent of the 9-year-olds in 1983-84 could read at the Basic level but only 17 percent could read at the Intermediate level are possible. The individual differences among students can be described without introducing the concepts of variance and standard deviation. Several tables containing levels of proficiency for NAEP students are shown in Part III.

* * *

In his early work on the measurement of learning outcomes, Glaser (1963) wrote:

> ...a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. (p. 519)

In this sense, the NAEP reading exercises have been used to create a criterion-referenced test. For the reading scale, we present not only sample exercises to show what students can do but also generalize to classes of behaviors.

This way of presenting results does not contain all of the information that presenting the percent passing for each item does but puts together a large amount of information in a simple way. The percents passing are

Figure 10.5-1
Levels of Proficiency

Rudimentary (150)

Readers who have acquired rudimentary reading skills and strategies can follow brief written directions. They can also select words, phrases, or sentences to describe a simple picture and can interpret simple written clues to identify a common object. Performance at this level suggests the ability to carry out simple, discrete reading tasks.

Basic (200)

Readers who have learned basic comprehension skills and strategies can locate and identify facts from simple informational paragraphs, stories, and news articles. In addition, they can combine ideas and make inferences based on short, uncomplicated passages. Performance at this level suggests the ability to understand specific or sequentially related information.

Intermediate (250)

Readers with the ability to use intermediate skills and strategies can search for, locate, and organize the information they find in relatively lengthy passages and can recognize paraphrases of what they have read. They can also make inferences and reach generalizations about main ideas and author's purpose from passages dealing with literature, science, and social studies. Performance at this level suggests the ability to search for specific information, interrelate ideas, and make generalizations.

Adept (300)

Readers with adept reading comprehension skills and strategies can understand complicated literary and informational passages, including material about topics they study at school. They can also analyze and integrate less familiar material and provide reactions to and explanations of the text as a whole. Performance at this level suggests the ability to find, understand, summarize, and explain relatively complicated information.

Advanced (350)

Readers who use advanced reading skills and strategies can extend and restructure the ideas presented in specialized and complex texts. Examples include scientific materials, literary essays, historical documents, and materials similar to those found in professional and technical working environments. They are also able to understand the links between ideas even when those links are not explicitly stated and to make appropriate generalizations even when the texts lack clear introductions or explanations. Performance at this level suggests the ability to synthesize and learn from specialized reading materials.

still available for interested researchers as long as they do not reveal the exercises to the public.

The anchoring of scales is not necessarily derivative from the IRT process, although the IRT parameters were used for NAEP. In fact, it is possible to attempt anchoring using any scale scores that are assigned to the students. Whether or not an item meets the criteria could be established directly by computing the percent correct at or above each scale point for each item. Items that met the criteria, if any, could be subjected to interpretation. Using the IRT parameter estimates is actually more conservative than necessary since we used the theoretical points at which 50 percent and 80 percent passed the item for evaluating an item's discrimination. If the IRT model holds, far more than 80 percent at much higher scores would pass the item and fewer than 50 percent of those scoring below the next lower level would pass the item. Scale anchoring is, therefore, applicable to any approximately unidimensional examination with highly discriminating items.

We should also note that there are other ways of anchoring the scale. For NAEP, we started with scale points and searched for general descriptions, words, and items to describe the scale points. This could be described as an extension of the suggestion of Bock, Mislevy, and Woodson (1982) to label points on a scale by items which 80 percent of the students at particular scale points could do. Another approach would be to start with behavioral descriptions and then look for the point on the scale which corresponded to the description; for example, we might select several exercises to represent a level of proficiency and then use some average measure of item difficulty as the point on the scale representing that proficiency level.

We also note that the percent above a scale point is not affected by monotone transformations of the theta scale. Once the anchor points are selected, any monotone transformation of the scale accompanied by the corresponding transformation of the anchor points will not affect the percentages at or above particular points (see Goldstein, 1980 for a clear description of the problem).

Chapter 11

THE WRITING DATA ANALYSIS: INTRODUCTION

Albert E. Beaton

Educational Testing Service

The NAEP has completed two reports on writing: *Writing Trends Across the Decade, 1974-84* (Applebee, Langer, & Mullis, 1986a) and *The Writing Report Card: Writing Achievement in American Schools, 1984* (Applebee, Langer, & Mullis, 1986b). The purpose of this chapter of the technical report is to provide the information necessary to understand the properties of the writing data, which are available on the public-use data tapes, and to understand the analyses underlying these two reports.

As mentioned in Chapter 1, ETS did not propose to include the writing assessment in the spiralled part of the sample nor did it intend to scale the writing data, but did indeed do both. The original conception was to present the trend data exercise by exercise using the tape-administered assessment results or, if the exercises administered by print could be reasonably equated with those administered by tape recorder, present a combination of the two data sets. The exercise-by-exercise approach was taken in the analysis for *Writing: Trends Across the Decade, 1974-84*.

Although seemingly simple at first, the exercise-by-exercise approach leads to complications in interpretation. With the NAEP data, a comparison of 17-year-olds of 1983-84 with their peers of 1978-79 would have to be based on different exercises than a comparison of the 17-year-olds with the 13-year-olds of 1983-84. Within the exercises used in a comparison, different exercise averages might move in the same direction, but at different rates. Different exercises have different averages; the reader who does not remember, for example, that the 13-year-olds had an easier exercise than the 17-year-olds may make false generalizations from the data.

We believed that a common scale onto which all writing exercises could be projected would help in the interpretation of the data. We strove to develop an overall measure of writing proficiency that would be comparable over ages and times.

Developing the writing scale incorporates much of the individual writing exercise information. Information at the exercise level is in *The Writing Report Card* and is available on the public-use data tapes for anyone interested in further research.

The development of the writing scale was not simple. First, we found that changing from tape recorded to printed administration affected the responses in ways that precluded equating the results; thus, the trend analyses were based only on data collected by tape recorded administration. We attempted to apply two IRT models for non-binary data that were proposed by Masters (1982), but these models did not provide acceptable results for the NAEP data. Finally, we developed the Average Response Method (ARM) (described below) and applied it to the cross-sectional analyses.

* * *

Before describing the data analyses, it is useful to review the background of the writing assessment data. In the 1983-84 writing assessment, NAEP used 22 different writing exercises. Exercises were designed to assess three different types of skills: informational writing, persuasive writing, and imaginative writing. The process by which these exercises were developed is described in Chapter 3.

NAEP has conducted four assessments of writing. Writing was first assessed in assessment Year 1 (1969-70), then in Year 5 (1973-74), and Year 10 (1978-79), and finally in Year 15 (1983-84). The NAEP writing exercises were supplied to the Educational Testing Service by the Education Commission of the States, which had administered the previous three writing assessments. ETS selected the 22 exercises that were used in Year 15. Some of these exercises had been used in previous writing assessments and others had not. No Year 1 exercises were used, because all of those exercises had been previously released.

The scoring of these exercises is discussed in Chapter 8.2. All exercises were scored using the primary trait method and a few were also scored using the holistic method. Some were also scored on secondary traits. All scores for the Year 15 assessment, including rescores for reliability analysis, are available on the public-use data tapes. The actual assessment papers were recovered for those students who had been assessed in NAEP Years 5 and 10 and had been administered essays that were also administered in Year 15. These papers were then scored along with the essays written for the 1983-84 assessment.

The changes in the design of NAEP by ETS have had an important effect on the data collected. In the earlier writing assessments, the writing exercises were administered using a tape recorder so that students were instructed about tasks aurally. The purpose of the tape recorder was to reduce dependence of subject area assessments on a student's ability to read an exercise and its instructions. The ETS design of NAEP led to administering different exercises, perhaps in different subject areas, at the same assessment session; thus, aural administration was not feasible. To avoid losing comparability between the Year 15 and prior writing assessments, two distinct types of assessments were performed, one using pencil-and-paper administration and the other using tape recorders. All of the 22 writing exercises were administered using pencil-and-paper methods; some of these were also administered using the tape recorders.

The pencil-and-paper assessment booklets were constructed using BIB spiralling where possible. The BIB spiralling generated 57 different assessment booklets at each grade/age level. The BIB-spiralled section of the pencil-and-paper assessment assures that each pair of exercises occurs jointly in some booklet and will be administered to an equivalent sample of students. (BIB spiralling is described in Chapter 4.) However, some reading and writing exercises took more than fourteen minutes to complete and thus could not fit into the fourteen-minute BIB block structure. To accommodate these exercises, four additional assessment booklets were developed at each grade/age level. These booklets, which are called the unbalanced incomplete block (UBIB) booklets, were spiralled into the pencil-and-paper sample and thus administered to a sample of students equivalent to the sample that received the BIB booklets. The UBIB booklets lose the property of having each exercise paired with each other exercise; in fact, few correlations are computable between exercises in different UBIB booklets.

Another important detail in understanding the writing data and their analysis is that the sample administered by tape recorder is collected only by age; the BIB and UBIB samples are collected by both age and grade. NAEP had sampled only by age in the past; thus, the part of the Year 15 assessment that was to be directly compared to past assessments was sampled in the same way. The BIB and UBIB samples contain students who are either 9-years-old or in the fourth grade, either 13-years-old or in the eighth grade, and either 17-years-old or in the eleventh grade. Since only age-eligible students were assessed using tape recorders, only age-eligible students were used to compare methods of administration.

Table 11(1) summarizes the properties of the writing data. For each exercise, the table provides:

- the exercise identification number and a short description of the exercise;
- the type of writing task;
- the assessment years in which the exercise was administered;
- an indicator of whether the exercise was in the BIB spiral, UBIB spiral, or paced tape samples;
- the identification number of the holistic score, if the exercise was scored holistically as well as by primary trait; and
- an indicator as to whether the exercise was used in computing ARM scale values. (The ARM scale is described in Chapter 11.4.)

Table 11(1)

Year 15 NAEP Writing Exercises

		Age 9 Assessment Year[1]			Age 13 Assessment Year			Age 17 Assessment Year			
Exercise	**Tasks[2]**	5	10	15	5	10	15	5	10	15	**Holistic**
N000102 DALI[3]	1		T[4]	B,T		T	B,T		T	B	N000108
N000202 SCHOOL RULE[3]	2			B			B			B	
N000302 RECREATION OPP.[3]	2						B			B	
N000402 FOOD ON FRONTIER[3]	1			B			B			B	
N000502 DISSECTING FROGS[3]	2						B				
N000602 XYZ COMPANY[3]	1			B			B				
N000702 SWIMMING POOL[3]	2			B			B			B	
N000802 PETS[3]	1		T	B		T	B				
N000902 RADIO STATION[3]	2			B			B			B	
N001002 APPLEBY HOUSE[3]	1			B			B			B	
N007202 HOLE IN THE BOX	3	T	T	U,T	T	T	U,T	T	T	U,T	N007208
N007602 FLASHLIGHT	3			U			U			U	
N007702 GHOST STORY	3			U			U			U	
N007902 FAVORITE MUSIC	1			U			U			U	
N008002 SPLIT SESSION	2					T	U,T		T	U,T	N008008
N014702 PLANTS	1			B							
N014802 SPACESHIP	2			B							
N014902 AUNT MAY	2		T	U,T							N014909
N018002 SPACE PROGRAM	2									B	
N019002 JOB APPLICATION	1									B	
N020002 UNCLE	2								T	B	
N021002 BIKE LANE	2									B	

[1] Year 5=1973-74, Year 10=1978-79, Year 15=1983-84

[2] Types of writing tasks: 1=informative, 2=persuasive, 3=imaginative

[3] Included in the ARM

[4] T=administered by tape recorder, age data only; B=administered in BIB spiral blocks, age and grade data; U=administered in other blocks, age and grade data

Much more information about the exercises is available in Chapter 3. The actual exercise text is available on the microfiche accompanying the public-use data tapes.

In addition to the responses to the writing exercises, analyses for the writing reports also include a number of specific questions about students' attitudes toward writing and their writing practices. A brief discussion of these items is included in Chapter 6; they are discussed more fully in the reports in which they are used.

The next four chapters of this technical report are summarized below.

11.1 The Writing Exercise Data

This section contains, among other things, the average values and standard deviations of the writing exercises and the inter-rater reliability coefficients.

11.2 The Effect of Mode of Exercise Administration (BIB Spiral or Paced Tape) on Estimates of Writing Performance

This section shows the differences in responses between the sample administered by pencil-and-paper and that administered by tape-recorded procedures. The comparison shows better average performance at all three age levels when the exercises are administered by tape recorder. The benefit attributable to tape-recorded administration appears to vary both by demographic subgroup and writing exercise. As a result, it was decided not to attempt to merge the data collected by the two methods.

11.3 Estimation of Trends in Writing Achievement

Because the amount of trend data was insufficient to support scaling, and because the pencil-and-paper data could not be merged with the data collected at tape recorder sessions, the analysis of trend data was based only on individual essays that were administered in different assessment years and which were also administered by tape recorder in Year 15. The statistical considerations in the trend analysis are discussed in this section.

11.4 The Average Response Method (ARM) of Scaling

Some of the writing data were scaled using the average response method. The underlying assumptions and derivations as well as the computation of plausible values are presented in this section. The potential bias due to model mis-specification is also discussed and an alternative method is given which is unbiased, but not as general in application. The two statistical procedures are compared using the NAEP writing data and the results are presented. It is our opinion that the ARM is a useful tool for estimation and interpretation and is a promising tool for future data analytic work.

Chapter 11.1

THE WRITING EXERCISE DATA

Albert E. Beaton

Educational Testing Service

The purpose of this section is to provide some basic information about the writing data. All data were rated by professional judges and the details of the scoring process are given in Chapter 8.2. The same scoring protocols were used for all three grade/age combinations and were applied to the data from past assessments as well. Included here is information about:

- the rater reliability;
- the scale drift during the rating process; and
- the basic descriptive statistics for each exercise.

Other information about the writing data can be found in Chapters 11.2, 11.3, and 11.4.

11.1.1 Inter-Rater Reliability

Since the individual essays were rated by professional judges, the question of the consistency of judges must be addressed. To do so, we performed an analysis of the inter-rater reliability. A 20 percent sample of the essays was selected and independently rated by a second scorer. These multiply-rated essays form the basis of the inter-rater reliability analysis, the results of which are shown in Table 11.1(1).

Two statistics were chosen for presentation: the percent of exact agreement and the reliability coefficient. The percent of exact agreement is the percentage of times that the two scorers agreed exactly in their ratings. The reliability coefficient is the intra-class correlation among raters.

The results for both primary trait and holistic scorings are shown in Table 11.1(1). For each grade/age combination, the number of responses analyzed is shown. The next column is the number of times the two scores agreed exactly in their ratings. The third column is the reliability coefficient.

Table 11.1(1)

Percentages of Exact Score Point Agreement and Intra-class Correlation Coefficients for Primary Trait Scoring, Year 15 (Possible Score Range: 0 to 4)

	Grade 4			Grade 8			Grade 11		
Writing Tasks	N	Agreement	Coefcnt.	N	Agreement	Coefcnt.	N	Agreement	Coefcnt.
Informative Writing									
Pets	534	92.3	.88	524	84.4	.78	0	-	-
Job Application	0	-	-	0	-	-	497	91.1	.92
Plants	402	92.1	.93	0	-	-	0	-	-
Appleby House	635	89.6	.92	719	79.0	.84	715	89.4	.92
XYZ Company	506	93.1	.92	466	89.9	.86	0	-	-
Dali	396	90.9	.88	468	82.0	.81	449	91.3	.92
Favorite Music	434	93.4	.89	528	84.4	.67	499	95.0	.90
Food on the Frontier	440	92.5	.89	460	82.2	.76	487	92.6	.90
Persuasive Writing									
School Rule	479	91.6	.88	483	81.4	.70	527	92.5	.91
Dissecting Frogs	0	-	-	466	78.0	.71	-	-	-
Swimming Pool	535	90.8	.89	523	83.9	.82	523	90.9	.91
Split Sessions	0	-	-	432	84.4	.80	461	88.4	.88
Space Ship	506	88.1	.90	0	-	-	-	-	-
Space Program	0	-	-	0	-	-	495	90.2	.92
Recreation Opportunity	0	-	-	452	86.4	.87	478	89.9	.92
Radio Station	639	95.7	.97	720	84.2	.88	-	-	-
Aunt May	434	91.6	.92	-	-	-	-	-	-
Uncle	0	-	-	-	-	-	523	89.3	.90
Bike Lane	0	-	-	-	-	-	720	88.5	.91
Imaginative Writing									
Hole in the Box	424	91.5	.89	461	82.6	.86	504	91.1	.92
Flashlight	445	92.9	.91	436	80.9	.79	463	92.3	.91
Ghost Story	435	93.3	.89	528	83.1	.85	498	91.1	.93

These results show a very high degree of agreement between the raters. Table 11.1(2) summarizes the statistics by grade.

For Grades 4 and 11, no exercise had less than 88 percent exact agreement; some exercises had agreement over 95 percent. The reliability coefficients are also high, ranging from .88 to .97.

The reliability for Grade 8 is quite acceptable, but not as high. Percents of exact agreement range from 78.0 to 89.9 and reliability coefficients from .67 to .88. The lower values for Grade 8 may be related to the fact that because the eighth graders were assessed first, in the fall of 1983, the scorers were less experienced when these papers were rated.

Table 11.1(3) shows the percents of exact agreement and reliability coefficients for the exercises that were used in the trend report. These essays were scored for the first time to estimate trends and are, necessarily, reported by ages, not grades, since past data were collected only by age. These results are also quite good.

11.1.2 Batching Effect

As mentioned above, the writing samples were rated as they were received by the scorers with the result that the eighth graders were rated first, the fourth graders next, and the eleventh graders last. Since there were so many essays to score, waiting until all writing samples were collected and then intermingling them, so that all grades would be rated at the same time, would have caused serious delays in reporting the writing results.

The rating of the different grades separately and serially led to a concern about a drift in the rating scale throughout the rating process. To examine the size of the drift, if one existed, an experiment on the effect of batching was performed.

The experiment was designed and carried out by Zwick. In summary, three essays which were administered in all three grades were selected. These three essays were contained in one booklet. Half of the booklets were retrieved, with resulting sample sizes of 156 cases for Grade 4/Age 9, 174 cases for Grade 8/Age 13, and 173 cases for Grade 11/Age 17.

These booklets were randomly permuted and then blindly re-rated; that is, the re-raters were given neither the age or grade of a respondent nor the previous rating of an exercise. The re-raters were selected from the pool of original raters. After the re-rating, the original rating and the re-rating were compared using a three-way (Grade/Age x Exercises x Time) repeated measures analysis of variance. It was decided before the analysis that a batch effect of less than a tenth of a score point was ignorable. The estimated batch effects were .01 for Grade 4/Age 9, -.04 for Grade 8/Age 13, and .03 for Grade 11/Age 17. These batch effects were not statistically significant at the .05 level.

Table 11.1(2)

Reliability Statistics for Primary Trait Ratings

Grade	Number of Exercises*	Low Percent	High Percent	Low r	High r
4	15	88.1	95.7	.88	.97
8	15	78.0	89.9	.67	.88
11	15	88.4	95.0	.88	.93

*Although there were 22 writing exercises over all grades, only 15 were administered at each grade.

Table 11.1(3)

Percentages of Exact Score Point Agreement and Intra-class Correlation Coefficients for Primary Trait Scoring Conducted in 1983-84

	1974 Papers			1979 Papers			1984 Papers		
	N	Agreement	Coefficient	N	Agreement	Coefficient	N	Agreement	Coefficient
Age 9									
Hole in the Box	501	92%	.90	497	93%	.89	289	90%	.86
Dali	0	--	--	509	88	.83	283	90	.83
Aunt May	0	--	--	512	88	.89	283	92	.95
Age 13									
Hole in the Box	505	85	.82	563	85	.83	282	78	.79
Dali	0	--	--	535	90	.86	274	78	.73
Split Sessions	0	--	--	574	90	.84	275	87	.79
Age 17									
Hole in the Box	459	90	.90	547	89	.89	332	92	.91
Dali	0	--	--	501	90	.85	337	90	.89
Split Sessions	0	--	--	555	91	.89	335	89	.91

More detail about, and other analyses of, the data collected for this experiment are provided in a supplementary paper by Zwick (1986b).

As a result of this experiment, it was decided to use the original scorings without any adjustment for batching.

11.1.3 Descriptive Statistics

Table 11.1(4) contains the number of students who responded to each writing exercise as well as the mean and standard deviation of the ratings. These statistics are presented for the three grade samples only. The sampling weights were used in calculating the means and standard deviations.

Table 11.1(4)

Number of Students Responding to Each Writing Exercise with Rating Mean and Standard Deviation (Possible Score Range: 0 to 4)

	Grade 4			Grade 8			Grade 11		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
N000102	1810	1.3610	0.6776	1970	1.8949	0.7267	2268	2.1632	0.7768
N000202	2018	1.6363	0.6029	2253	1.9808	0.5926	2370	2.1320	0.6442
N000302	0	-	-	2234	1.6082	0.7751	2357	1.9565	0.8116
N000402	1844	1.3751	0.6095	2236	2.0015	0.6680	2373	2.1064	0.6974
N000502	0	-	-	2339	2.0276	0.6260	0	-	-
N000602	1770	1.7714	0.9676	2229	2.5226	0.7474	0	-	-
N000702	2027	1.5156	0.6911	2341	1.7850	0.6864	2400	1.9511	0.7274
N000802	1698	1.7073	0.5960	2190	2.1536	0.7259	0	-	-
N000902	2066	1.5686	0.8433	2305	2.0762	0.8871	0	-	-
N001002	1497	1.8914	0.8248	2040	2.4325	0.8015	2050	2.4953	0.8320
N007202	2139	1.3386	0.6956	2294	1.7718	0.8994	2469	1.8035	0.8463
N007602	2018	1.7443	0.6418	2286	2.2248	0.6998	2362	2.2833	0.7081
N007702	2119	1.8467	0.6496	2336	2.2505	0.7986	2429	2.3175	0.9324
N007902	1564	1.5294	0.5760	1990	1.8863	0.5335	2127	1.8761	0.5272
N008002	0	-	-	2330	1.4060	0.7146	2376	1.7082	0.7948
N014702	2029	2.2458	0.7478	0	-	-	0	-	-
N014802	2026	1.8456	0.8709	0	-	-	0	-	-
N014902	2102	1.6874	0.9646	0	-	-	0	-	-
N018002	0	-	-	0	-	-	2440	2.0740	0.8368
N019002	0	-	-	0	-	-	2325	2.4733	0.9226
N020002	0	-	-	0	-	-	2156	1.9478	0.8013
N021002	0	-	-	0	-	-	2433	1.9333	0.8653
N000108	2004	2.6170	1.3651	2266	2.9622	1.2648	2379	3.4443	1.3081
N007208	2138	2.5355	1.4528	2294	2.7952	1.5562	2469	3.2046	1.5125
N008008	0	-	-	2328	2.819	1.2713	2376	3.3319	1.3617
N014909	2103	2.5873	1.3299	0	-	-	0	-	-

Chapter 11.2

THE EFFECT OF MODE OF ITEM ADMINISTRATION (BIB SPIRAL OR PACED TAPE) ON ESTIMATES OF WRITING PERFORMANCE[1]

Eugene G. Johnson

Educational Testing Service

The Year 15 NAEP writing assessment, the fourth such assessment in the history of NAEP, is the first writing assessment in which the Balanced Incomplete Block spiral design was used for assigning exercises to students. In the three earlier writing assessments, Year 1 (1969-70), Year 5 (1973-74) and Year 10 (1978-79), the total battery of writing items was divided into a number of mutually exclusive booklets, called packages, and each such package was, in turn, administered to a nationally representative sample of students. While this matrix design allows analysis of the interrelationships between exercises appearing in the same package, the interrelationships between exercises in different packages cannot be readily estimated, because no student was administered more than one of the packages.

The Year 15 NAEP design has remedied this deficiency through a complex variant of matrix sampling called balanced incomplete block (BIB) spiralling. Details of this procedure appear in Chapter 5. In brief, the total assessment battery (of both reading and writing exercises) was divided into item blocks requiring an assessment time of fourteen minutes. Each of these blocks was then assigned to 57 assessment booklets in such a manner that each booklet consisted of three blocks and each block of exercises was paired with every other block in at least one of the booklets. Since some writing items required a response time longer than the fourteen minutes permitted in the BIB design, six special booklets, the unbalanced incomplete block (UBIB) booklets, were created to accommodate these items. Each UBIB booklet consisted of one "double block", containing a longer item and requiring 30 minutes of testing time, and one of the regular BIB blocks.

The total set of BIB and UBIB booklets were then spiralled, cycling the booklets for administration so that, typically, no two students in any assessment session in a school received the same booklet. More importantly, every item block and every pair of item blocks (within the BIB portion of the assessment) was administered to a representative sample of students, enabling the examination of interrelationships between all items encompassed by the BIB blocks. (For UBIB booklets, interrelationships can only be directly estimated for certain of the items).

---

[1] The statistical programming for this section was provided by Bruce Kaplan. The figures were produced by Ira Sample.

The change to the BIB spiralling design results in improved sampling efficiency and analysis potential, but at a cost. Prior to the Year 15 assessment, assessments of writing (and all other areas) were accompanied by paced audiotapes of the exercise stimuli. The advantage of such a mode of administration is that it allows for the separation of reading ability from the subject area being assessed. In paced administrations of the writing assessment, the instructions for the exercise are read aloud so students can respond to the exercise even though they may have difficulty reading the instructions. This type of administration was possible because all students in a particular paced assessment session received the same package. Because each student in a BIB spiralled assessment session has typically received a different booklet, it is not possible to accompany a BIB spiralled assessment session with paced audiotapes.

To determine the effect of this change in mode of administration (from paced to BIB spiralled) on estimates of writing achievement, a selected subset of writing exercises was administered both as part of the primary BIB spiralled assessment and as part of a much smaller paced tape assessment. The Year 15 paced tape assessment was also designed to ascertain the effect of change in mode of administration on estimates of reading achievement (the results of this are reported in Section 10.3.6). This portion of the Year 15 assessment of reading and writing was based on an additional administration of approximately one third of the reading and writing exercises by the previously used paced tape procedures.

The exercises to be administered by paced tape procedures at a given age were divided into four distinct packages. Each package was then administered to a probability sample of students representative of the nation. Between 1,300 and 1,600 students responded to each of these packages. Each of the paced tape packages was administered in exactly the same manner as the paced administrations in past assessments.

Because writing exercises generally require more response time than reading exercises, fewer writing exercises could be chosen as a part of the BIB-pace comparison. Three writing exercises were chosen for this purpose at each assessment age level. The criteria for selection were:

(1) The exercises had to have been administered in the previous (Year 10) writing assessment and, if possible, also in the Year 5 assessment.

(2) The exercises were to be representative of each of the three major purposes of writing as measured by the informative, persuasive and imaginative tasks.

(3) Subject to 1 and 2, each of the exercises was to be given to more than one age.

The result of this selection are the four writing exercises shown in Table 11.2(1). Of these four exercises, two were assigned to all three ages ("Hole in the Box" and "Dali"). One of the remaining two, "Split Session", had been given at ages 13 and 17 only and was replaced by "Aunt May" for the age 9 comparison. One exercise, "Hole in the Box", an imaginative task, was presented in both the Year 5 and Year 10 assessments. The remaining three exercises, the informative task "Dali" and the persuasive tasks "Split Session" and "Aunt May", were previously presented only in the Year 10 assessment.

The assignment of the exercises to the paced tape packages is also shown in Table 11.2(1). Of the four packages administered at a given age, two included writing exercises. One of these packages, P2, included two writing exercises--"Dali" (at Age 9) and either "Aunt May" or "Split Session" (at Ages 13 and 17). Consequently, the estimates for these exercises are based on the same sample of students in a given age. The remaining package, P4, included a single exercise, "Hole in the Box"; estimates for this exercise for an age are based on a different, but randomly equivalent, subsample of students.

The responses to the exercises from both modes of administration were professionally scored for task accomplishment (primary trait scoring). (A discussion of professional scoring is provided in Chapter 8.2.) The five levels of proficiency used to categorize the responses, along with their numeric codes, are:

0: unrateable
1: unsatisfactory
2: minimal
3: adequate
4: elaborated

Assessment results are reported both in terms of the proportion of students whose writing reaches or exceeds a given level of proficiency and in terms of mean proficiency levels.

Tables 11.2(2) through 11.2(10) show the comparison of the estimates of writing proficiency for the BIB and paced modes of administration by age and for a selected set of demographic subgroups within each age. Each table shows both the estimated percent of students of a given type scoring at or above the minimal (2) proficiency level and the estimated mean proficiency level for the subgroup. The numbers in parentheses are the estimates of the sampling standard errors of these proficiency estimates. Also included are the differences in proficiency level between the BIB and paced modes of administration (DIFFER), accompanied by a standard error.

Figures 11.2-1, 11.2-2 and 11.2-3 are plots by subgroup and age of the differences between the percent at or above minimum proficiency. In general, the previous writing assessment procedures using paced audiotapes are significantly less difficult for a student than the BIB spiralled procedure, which relies on a student's ability to read and understand the instructions given by the writing prompt. In every case where there is a significant difference, responses were rated better for the paced mode of administration.

Furthermore, the effect of mode of administration is differential in that differences in performance levels are greater for some subgroups than for others. The effect of mode of administration also varies from item to item within subgroup.

Because of the differential effect of mode of administration across items and subgroups, it was felt that the responses to the BIB and paced modes of administration could not be reliably equated. This has important consequences in the measurement of trends in writing achievement across time. These consequences are discussed in the following chapter.

Table 11.2(1)

Writing Exercises Selected for the BIB/Pace Comparison

Exercise	Task	Ages	Assessment Years	Package
Hole in the Box	Imaginative	9, 13, 17	Years 5, 10, 15	P4
Dali	Informative	9, 13, 17	Years 10, 15	P2
Aunt May	Persuasive	9	Years 10, 15	P2
Split Session	Persuasive	13, 17	Years 10, 15	P2

Table 11.2(2)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Aunt May"

		N	MEAN	% >= 2
-- TOTAL --	BIB	1960	1.61(0.03)	44.99(1.30)
	PACED	1356	1.90(0.04)	58.24(2.02)
	DIFFER		-0.29(0.05)*	-13.24(2.40)*
ETHNICITY/RACE				
White	BIB	1316	1.71(0.03)	49.75(1.36)
	PACED	869	2.01(0.05)	63.64(2.34)
	DIFFER		-0.30(0.06)*	-13.89(2.70)*
Black	BIB	276	1.24(0.06)	28.30(2.92)
	PACED	223	1.57(0.06)	43.69(3.69)
	DIFFER		-0.33(0.09)*	-15.39(4.70)*
Hispanic	BIB	288	1.34(0.05)	33.24(3.20)
	PACED	203	1.55(0.10)	40.93(5.36)
	DIFFER		-0.21(0.12)	-7.69(6.24)
PARENTAL EDUCATION				
Not graduated H.S.	BIB	125	1.41(0.07)	40.46(5.42)
	PACED	76	1.68(0.10)	45.72(5.52)
	DIFFER		-0.27(0.12)*	-5.26(7.73)
Graduated H.S.	BIB	378	1.62(0.05)	43.36(2.92)
	PACED	280	1.88(0.06)	54.96(3.02)
	DIFFER		-0.26(0.08)*	-11.60(4.20)*
Post H.S.	BIB	715	1.76(0.05)	52.83(2.39)
	PACED	472	2.01(0.05)	63.89(2.34)
	DIFFER		-0.24(0.07)*	-11.06(3.34)*
Unknown	BIB	713	1.48(0.04)	39.19(2.09)
	PACED	514	1.85(0.06)	57.21(3.05)
	DIFFER		-0.37(0.07)*	-18.01(3.70)*

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(2)
(continued)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Aunt May"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	246	1.27(0.05)	31.56(2.37)
	!PACED	194	1.67(0.07)	46.94(2.90)
	DIFFER		-0.39(0.09)*	-15.39(3.74)*
Advantaged Urban	BIB	247	1.85(0.09)	55.51(4.47)
	!PACED	183	2.27(0.07)	74.79(2.33)
	DIFFER		-0.42(0.12)*	-19.27(5.04)*
GRADE				
< Modal Grade	BIB	576	1.26(0.05)	29.06(2.16)
	PACED	458	1.58(0.07)	44.42(3.92)
	DIFFER		-0.33(0.08)*	-15.36(4.48)*
At Modal Grade	BIB	1378	1.71(0.03)	49.84(1.40)
	PACED	893	2.05(0.05)	64.98(2.17)
	DIFFER		-0.34(0.06)*	-15.14(2.58)*
> Modal Grade	!BIB	6	2.09(0.77)	43.94(28.29)
	!PACED	5	2.22(0.37)	91.29(10.05)
	DIFFER		-0.13(0.85)	-47.35(30.02)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(3)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
-- TOTAL --	BIB	1680	1.32(0.02)	39.12(1.38)
	PACED	1356	1.57(0.03)	55.46(1.89)
	DIFFER		-0.24(0.03)*	-16.33(2.34)*
ETHNICITY/RACE				
White	BIB	1132	1.39(0.02)	43.52(1.70)
	PACED	869	1.62(0.03)	59.58(2.45)
	DIFFER		-0.23(0.04)*	-16.06(2.98)*
Black	BIB	243	1.08(0.05)	24.79(3.38)
	PACED	223	1.35(0.05)	38.64(4.41)
	DIFFER		-0.27(0.07)*	-13.86(5.55)*
Hispanic	BIB	227	1.21(0.06)	31.16(4.55)
	PACED	203	1.44(0.05)	46.57(3.87)
	DIFFER		-0.24(0.08)*	-15.42(5.97)*
PARENTAL EDUCATION				
Not graduated H.S.	BIB	95	1.14(0.07)	27.68(4.78)
	PACED	76	1.51(0.07)	51.65(6.05)
	DIFFER		-0.37(0.10)*	-23.98(7.71)*
Graduated H.S.	BIB	332	1.19(0.04)	30.89(2.87)
	PACED	280	1.53(0.04)	54.37(3.94)
	DIFFER		-0.34(0.06)*	-23.47(4.87)*
Post H.S.	BIB	622	1.45(0.03)	47.85(2.37)
	PACED	472	1.65(0.05)	61.04(2.91)
	DIFFER		-0.19(0.06)*	-13.20(3.76)*
Unknown	BIB	615	1.29(0.03)	36.78(2.06)
	PACED	514	1.53(0.03)	52.53(2.55)
	DIFFER		-0.24(0.04)*	-15.75(3.28)*

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(3)
(continued)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	202	1.17(0.06)	26.20(4.35)
	!PACED	194	1.40(0.06)	41.56(5.38)
	DIFFER		-0.23(0.09)*	-15.37(6.92)
Advantaged Urban	BIB	218	1.48(0.07)	47.95(3.27)
	!PACED	183	1.73(0.04)	68.21(3.33)
	DIFFER		-0.24(0.08)*	-20.26(4.67)*
GRADE				
< Modal Grade	BIB	459	1.06(0.03)	20.56(2.27)
	PACED	458	1.36(0.04)	38.21(3.11)
	DIFFER		-0.30(0.05)*	-17.64(3.85)*
At Modal Grade	BIB	1214	1.39(0.03)	43.84(1.63)
	PACED	893	1.67(0.03)	64.01(2.03)
	DIFFER		-0.28(0.04)*	-20.17(2.60)*
> Modal Grade	!BIB	7	1.96(0.16)	84.60(11.23)
	!PACED	5	1.73(0.27)	72.56(27.23)
	DIFFER		0.24(0.32)	12.03(29.45)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(4)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2029	1.30(0.02)	37.21(1.59)
	PACED	1344	1.55(0.03)	54.56(1.97)
	DIFFER		-0.25(0.04)*	-17.34(2.53)*
ETHNICITY/RACE				
White	BIB	1345	1.34(0.03)	38.47(1.98)
	PACED	832	1.58(0.03)	56.56(2.21)
	DIFFER		-0.24(0.04)*	-18.09(2.96)*
Black	BIB	308	1.19(0.05)	34.28(2.69)
	PACED	178	1.45(0.09)	47.85(7.91)
	DIFFER		-0.25(0.10)*	-13.57(8.36)
Hispanic	BIB	273	1.22(0.07)	32.38(3.77)
	PACED	263	1.46(0.07)	47.95(4.38)
	DIFFER		-0.23(0.10)*	-15.57(5.78)*
PARENTAL EDUCATION				
Not graduated H.S.	BIB	121	1.21(0.09)	26.47(5.16)
	PACED	104	1.50(0.05)	48.03(4.30)
	DIFFER		-0.29(0.10)*	-21.56(6.71)*
Graduated H.S.	BIB	392	1.29(0.04)	33.55(2.25)
	PACED	277	1.47(0.05)	47.33(3.55)
	DIFFER		-0.18(0.06)*	-13.78(4.20)*
Post H.S.	BIB	724	1.40(0.04)	44.66(2.55)
	PACED	453	1.65(0.04)	62.24(2.78)
	DIFFER		-0.26(0.05)*	-17.58(3.77)*
Unknown	BIB	773	1.25(0.03)	34.35(2.51)
	PACED	495	1.52(0.04)	53.14(2.61)
	DIFFER		-0.28(0.05)*	-18.80(3.62)*

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(4)
(continued)

Effect of Mode of Administration on Writing Performance

Age 9 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	276	1.18(0.06)	35.58(4.03)
	!PACED	205	1.58(0.08)	60.06(4.15)
	DIFFER		-0.40(0.10)*	-24.47(5.78)*
Advantaged Urban	BIB	232	1.46(0.05)	48.13(4.62)
	!PACED	90	1.77(0.05)	71.26(3.03)
	DIFFER		-0.31(0.08)*	-23.14(5.53)*
GRADE				
< Modal Grade	BIB	635	1.06(0.04)	21.93(2.76)
	PACED	433	1.40(0.04)	45.27(2.78)
	DIFFER		-0.34(0.05)*	-23.34(3.91)*
At Modal Grade	BIB	1386	1.38(0.02)	42.08(1.72)
	PACED	907	1.62(0.03)	58.83(2.33)
	DIFFER		-0.24(0.04)*	-16.75(2.89)*
> Modal Grade	!BIB	8	1.77(0.26)	76.74(26.37)
	!PACED	4	2.00(0.00)	100.00(0.0)
	DIFFER		-0.23(0.26)	-23.26(26.37)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(5)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Split Session"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2241	1.37(0.02)	31.77(1.10)
	PACED	1276	1.43(0.02)	34.08(1.61)
	DIFFER		-0.06(0.03)	-2.31(1.95)
ETHNICITY/RACE				
White	BIB	1631	1.42(0.02)	34.59(1.13)
	PACED	889	1.48(0.03)	36.56(1.62)
	DIFFER		-0.06(0.03)	-1.98(1.98)
Black	BIB	293	1.26(0.06)	24.03(3.17)
	PACED	211	1.30(0.06)	28.12(4.37)
	DIFFER		-0.04(0.08)	-4.09(5.40)
Hispanic	BIB	264	1.18(0.04)	21.07(2.61)
	PACED	126	1.21(0.06)	20.41(6.70)
	DIFFER		-0.03(0.07)	0.67(7.19)
PARENTAL EDUCATION				
Not graduated H.S.	BIB	208	1.18(0.05)	22.52(3.58)
	PACED	92	1.23(0.05)	23.45(5.02)
	DIFFER		-0.05(0.07)	-0.93(6.17)
Graduated H.S.	BIB	784	1.32(0.03)	28.16(1.87)
	PACED	451	1.39(0.03)	31.67(2.39)
	DIFFER		-0.07(0.04)	-3.51(3.04)
Post H.S.	BIB	1003	1.49(0.03)	38.86(1.78)
	PACED	574	1.54(0.04)	40.53(2.10)
	DIFFER		-0.04(0.04)	-1.67(2.75)
Unknown	BIB	226	1.19(0.05)	22.16(3.23)
	PACED	130	1.17(0.06)	18.08(3.79)
	DIFFER		0.02(0.07)	4.08(4.98)

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(5)
(continued)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Split Session"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	229	1.21(0.06)	22.95(4.11)
	!PACED	141	1.20(0.12)	20.06(8.38)
	DIFFER		0.02(0.14)	2.89(9.34)
Advantaged Urban	!BIB	264	1.47(0.03)	37.34(2.43)
	!PACED	81	1.55(0.06)	42.02(4.86)
	DIFFER		-0.09(0.07)	-4.67(5.43)
GRADE				
< Modal Grade	BIB	655	1.18(0.03)	20.76(2.04)
	PACED	393	1.27(0.03)	24.52(2.11)
	DIFFER		-0.10(0.04)*	-3.76(2.94)
At Modal Grade	BIB	1579	1.46(0.02)	36.62(1.44)
	PACED	882	1.50(0.03)	38.37(2.01)
	DIFFER		-0.04(0.04)	-1.75(2.47)
> Modal Grade	!BIB	7	1.21(0.13)	20.93(13.22)
	!PACED	1	1.00(0.0)	0.0 (0.0)
	DIFFER		0.21(0.13)	20.93(13.22)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(6)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
-- TOTAL --	BIB	1890	1.90(0.02)	71.67(1.21)
	PACED	1276	2.01(0.02)	81.40(1.20)
	DIFFER		-0.12(0.03)*	-9.73(1.70)*
ETHNICITY/RACE				
White	BIB	1425	1.96(0.03)	75.36(1.55)
	PACED	889	2.09(0.03)	84.94(1.37)
	DIFFER		-0.14(0.04)*	-9.57(2.07)*
Black	BIB	216	1.61(0.05)	54.80(3.45)
	PACED	211	1.72(0.03)	67.94(3.42)
	DIFFER		-0.12(0.06)*	-13.14(4.86)*
Hispanic	BIB	195	1.68(0.06)	59.08(3.11)
	PACED	126	1.76(0.06)	72.33(4.75)
	DIFFER		-0.08(0.09)	-13.24(5.68)*
PARENTAL EDUCATION				
Not graduated H.S.	BIB	149	1.68(0.08)	62.60(5.73)
	PACED	92	1.75(0.06)	74.98(5.51)
	DIFFER		-0.08(0.10)	-12.38(7.95)
Graduated H.S.	BIB	659	1.81(0.03)	68.49(2.22)
	PACED	451	1.95(0.03)	80.29(1.83)
	DIFFER		-0.13(0.04)*	-11.80(2.88)*
Post H.S.	BIB	890	2.04(0.03)	77.70(1.43)
	PACED	574	2.12(0.03)	84.45(1.75)
	DIFFER		-0.09(0.04)*	-6.74(2.26)*
Unknown	BIB	177	1.65(0.04)	58.48(4.42)
	PACED	130	1.80(0.06)	74.39(4.62)
	DIFFER		-0.15(0.07)*	-15.90(6.39)*

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(6)
(continued)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	167	1.67(0.06)	58.83(4.19)
	!PACED	141	1.66(0.09)	63.56(7.64)
	DIFFER		0.01(0.11)	-4.73(8.72)
Advantaged Urban	!BIB	225	2.27(0.08)	87.10(2.36)
	!PACED	81	2.25(0.06)	87.74(2.28)
	DIFFER		0.01(0.11)	-0.64(3.28)
GRADE				
< Modal Grade	BIB	545	1.67(0.03)	59.96(2.20)
	PACED	393	1.85(0.04)	78.35(2.43)
	DIFFER		-0.17(0.05)*	-18.39(3.28)*
At Modal Grade	BIB	1336	1.99(0.02)	76.42(1.38)
	PACED	882	2.09(0.03)	82.74(1.55)
	DIFFER		-0.10(0.04)*	-6.32(2.07)*
> Modal Grade	!BIB	9	2.47(0.33)	95.48(4.89)
	!PACED	1	3.00(0.00)	100.00(0.0)
	DIFFER		-0.53(0.33)	-4.52(4.89)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(7)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2290	1.74(0.02)	60.31(1.43)
	PACED	1289	1.84(0.04)	66.68(2.15)
	DIFFER		-0.09(0.04)*	-6.37(2.58)*
ETHNICITY/RACE				
White	BIB	1640	1.81(0.02)	63.33(1.55)
	PACED	915	1.83(0.04)	65.73(2.35)
	DIFFER		-0.02(0.05)	-2.39(2.82)
Black	BIB	320	1.48(0.07)	48.65(3.89)
	PACED	160	1.92(0.06)	74.23(3.26)
	DIFFER		-0.43(0.09)*	-25.57(5.08)*
Hispanic	BIB	248	1.55(0.09)	52.69(4.65)
	PACED	178	1.73(0.06)	62.74(6.17)
	DIFFER		-0.18(0.10)	-10.04(7.73)
PARENTAL EDUCATION				
Not graduated H.S.	BIB	183	1.64(0.05)	55.62(3.61)
	PACED	114	1.85(0.10)	62.83(6.67)
	DIFFER		-0.21(0.11)	-7.21(7.58)
Graduated H.S.	BIB	831	1.72(0.03)	60.49(1.65)
	PACED	470	1.74(0.05)	63.06(3.13)
	DIFFER		-0.02(0.06)	-2.57(3.53)
Post H.S.	BIB	1012	1.85(0.03)	63.80(1.73)
	PACED	567	1.95(0.05)	72.19(2.31)
	DIFFER		-0.10(0.05)	-8.39(2.88)*
Unknown	BIB	233	1.42(0.07)	45.59(4.52)
	PACED	130	1.71(0.07)	59.28(4.21)
	DIFFER		-0.29(0.10)*	-13.69(6.18)

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(7)
(continued)

Effect of Mode of Administration on Writing Performance

Age 13 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	BIB	229	1.51(0.07)	44.81(4.24)
	!PACED	113	1.91(0.05)	67.26(4.24)
	DIFFER		-0.40(0.09)*	-22.45(5.99)*
Advantaged Urban	!BIB	254	2.16(0.08)	81.72(3.23)
	!PACED	123	2.04(0.18)	69.19(8.71)
	DIFFER		0.11(0.20)	12.53(9.29)
GRADE				
< Modal Grade	BIB	736	1.50(0.04)	49.06(2.67)
	PACED	431	1.72(0.06)	58.98(4.09)
	DIFFER		-0.21(0.07)*	-9.92(4.88)
At Modal Grade	BIB	1548	1.86(0.02)	65.89(1.65)
	PACED	854	1.89(0.04)	70.53(2.18)
	DIFFER		-0.04(0.05)	-4.64(2.73)
> Modal Grade	!BIB	6	1.93(0.68)	47.29(23.09)
	!PACED	4	2.71(0.19)	100.00(0.0)
	DIFFER		-0.78(0.71)	-52.71(23.09)*

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(8)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Split Session"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2382	1.71(0.01)	59.70(0.83)
	PACED	1540	1.82(0.04)	63.79(2.54)
	DIFFER		-0.11(0.04)*	-4.09(2.67)
ETHNICITY/RACE				
White	BIB	1705	1.77(0.02)	62.61(1.36)
	PACED	1079	1.87(0.05)	66.95(2.92)
	DIFFER		-0.10(0.05)	-4.34(3.22)
Black	BIB	370	1.48(0.04)	48.65(2.92)
	PACED	242	1.68(0.05)	54.73(3.15)
	DIFFER		-0.20(0.07)*	-6.09(4.29)
Hispanic	BIB	236	1.59(0.07)	53.49(5.07)
	PACED	163	1.68(0.08)	57.19(5.99)
	DIFFER		-0.09(0.10)	-3.70(7.85)
PARENTAL EDUCATION				
Not graduated H.S.	BIB	313	1.53(0.05)	48.63(3.67)
	PACED	194	1.74(0.05)	63.49(3.42)
	DIFFER		-0.22(0.07)*	-14.86(5.01)*
Graduated H.S.	BIB	833	1.71(0.02)	61.08(1.61)
	PACED	558	1.76(0.03)	61.38(2.13)
	DIFFER		-0.05(0.04)	-0.30(2.67)
Post H.S.	BIB	1146	1.78(0.02)	63.04(1.12)
	PACED	696	1.91(0.08)	67.21(4.44)
	DIFFER		-0.13(0.08)	-4.17(4.58)
Unknown	!BIB	74	1.28(0.08)	34.71(5.08)
	!PACED	54	1.29(0.08)	28.90(6.39)
	DIFFER		-0.01(0.11)	5.81(8.17)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated

Table 11.2(8)
(continued)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Split Session"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	!BIB	254	1.51(0.04)	52.99(3.11)
	!PACED	181	1.67(0.06)	55.73(3.15)
	DIFFER		-0.16(0.07)*	-2.74(4.42)
Advantaged Urban	BIB	289	1.79(0.05)	61.54(3.56)
	PACED	190	1.85(0.21)	62.31(13.19)
	DIFFER		-0.07(0.22)	-0.77(13.66)
GRADE				
< Modal Grade	BIB	439	1.53(0.03)	50.39(2.08)
	PACED	310	1.55(0.05)	49.55(4.06)
	DIFFER		-0.02(0.06)	0.84(4.56)
At Modal Grade	BIB	1748	1.74(0.02)	61.54(0.99)
	PACED	1097	1.89(0.05)	67.38(2.75)
	DIFFER		-0.15(0.05)*	-5.84(2.92)
> Modal Grade	BIB	195	1.89(0.08)	68.85(3.94)
	PACED	133	1.85(0.09)	65.09(5.63)
	DIFFER		0.04(0.11)	3.77(6.88)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(9)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2282	2.13(0.02)	81.95(1.05)
	PACED	1540	2.28(0.04)	88.99(1.19)
	DIFFER		-0.14(0.05)*	-7.04(1.58)*
ETHNICITY/RACE				
White	BIB	1706	2.21(0.03)	84.98(1.20)
	PACED	1079	2.35(0.05)	90.95(1.30)
	DIFFER		-0.14(0.05)*	-5.98(1.76)*
Black	BIB	284	1.82(0.05)	68.80(2.62)
	PACED	242	1.94(0.05)	79.71(2.99)
	DIFFER		-0.12(0.07)	-10.90(3.98)*
Hispanic	BIB	217	1.86(0.06)	73.87(3.50)
	PACED	163	2.17(0.07)	88.78(2.09)
	DIFFER		-0.31(0.09)*	-14.91(4.08)*
PARENTAL EDUCATION				
Not graduated H.S.	BIB	275	1.89(0.05)	75.32(3.10)
	PACED	194	2.16(0.07)	87.36(2.89)
	DIFFER		-0.27(0.08)*	-12.04(4.24)*
Graduated H.S.	BIB	799	2.00(0.03)	76.90(1.46)
	PACED	558	2.25(0.05)	90.02(1.83)
	DIFFER		-0.25(0.06)*	-13.12(2.34)*
Post H.S.	BIB	1116	2.30(0.04)	87.93(1.38)
	PACED	696	2.37(0.07)	90.21(1.68)
	DIFFER		-0.07(0.08)	-2.28(2.17)
Unknown	BIB	68	1.63(0.11)	63.90(5.19)
	PACED	54	1.78(0.13)	74.72(6.84)
	DIFFER		-0.15(0.17)	-10.82(8.59)

* Significant difference between BIB and Pace (Alpha = .05)

Table 11.2(9)
(continued)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Dali"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	!BIB	255	1.84(0.04)	70.94(3.03)
	!PACED	181	1.89(0.11)	74.00(3.99)
	DIFFER		-0.05(0.12)	-3.06(5.01)
Advantaged Urban	BIB	293	2.25(0.10)	85.47(3.74)
	PACED	190	2.36(0.15)	88.65(2.21)
	DIFFER		-0.11(0.18)	-3.18(4.35)
GRADE				
< Modal Grade	BIB	391	1.83(0.05)	69.67(2.45)
	PACED	310	2.04(0.06)	82.35(2.48)
	DIFFER		-0.20(0.08)*	-12.68(3.48)*
At Modal Grade	BIB	1703	2.19(0.03)	84.33(1.13)
	PACED	1097	2.34(0.05)	91.38(1.22)
	DIFFER		-0.15(0.05)*	-7.04(1.66)*
> Modal Grade	BIB	188	2.34(0.04)	91.56(1.45)
	PACED	133	2.24(0.08)	83.90(4.27)
	DIFFER		0.10(0.09)	7.66(4.51)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(10)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
-- TOTAL --	BIB	2416	1.81(0.03)	66.48(1.41)
	PACED	1534	1.98(0.04)	75.13(2.05)
	DIFFER		-0.17(0.05)*	-8.65(2.49)*
ETHNICITY/RACE				
White	BIB	1750	1.89(0.03)	69.87(1.70)
	PACED	1130	2.02(0.04)	76.57(2.27)
	DIFFER		-0.13(0.06)*	-6.70(2.84)*
Black	BIB	377	1.53(0.04)	54.95(2.16)
	PACED	193	1.88(0.08)	69.65(4.30)
	DIFFER		-0.35(0.09)*	-14.70(4.81)*
Hispanic	BIB	224	1.59(0.08)	57.06(4.21)
	PACED	172	1.88(0.10)	71.15(4.57)
	DIFFER		-0.29(0.12)*	-14.09(6.21)*
PARENTAL EDUCATION				
Not graduated H.S.	BIB	301	1.63(0.04)	60.72(2.71)
	PACED	161	1.92(0.08)	72.36(4.28)
	DIFFER		-0.29(0.09)*	-11.64(5.06)*
Graduated H.S.	BIB	853	1.74(0.03)	63.22(1.62)
	PACED	543	1.92(0.05)	72.43(2.56)
	DIFFER		-0.18(0.06)*	-9.21(3.03)*
Post H.S.	BIB	1148	1.95(0.04)	71.86(2.06)
	PACED	775	2.08(0.04)	79.44(1.86)
	DIFFER		-0.13(0.06)*	-7.58(2.78)*
Unknown	!BIB	87	1.26([illegible].08)	43.47(5.79)
	!PACED	52	1.33(0.15)	47.97(7.34)
	DIFFER		-0.08(0.17)	-4.50(9.35)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Table 11.2(10)
(continued)

Effect of Mode of Administration on Writing Performance

Age 17 Primary Trait Score - "Hole in the Box"

		N	MEAN	% >= 2
SIZE/TYPE OF COMMUNITY				
Disadvantaged Urban	!BIB	241	1.54(0.05)	56.84(3.49)
	!PACED	179	1.82(0.10)	72.02(4.83)
	DIFFER		-0.28(0.12)*	-15.18(5.95)*
Advantaged Urban	BIB	313	2.10(0.09)	78.12(3.21)
	!PACED	221	2.05(0.18)	74.30(8.32)
	DIFFER		0.05(0.20)	3.82(8.92)
GRADE				
< Modal Grade	BIB	427	1.57(0.05)	56.60(3.28)
	PACED	253	1.68(0.08)	60.69(4.47)
	DIFFER		-0.11(0.10)	-4.09(5.55)
At Modal Grade	BIB	1780	1.86(0.03)	68.87(1.53)
	PACED	1170	2.04(0.04)	77.49(2.07)
	DIFFER		-0.18(0.05)*	-8.62(2.58)*
> Modal Grade	BIB	209	2.03(0.08)	71.23(3.50)
	PACED	111	2.10(0.05)	82.38(3.70)
	DIFFER		-0.07(0.09)	-11.15(5.10)

* Significant difference between BIB and Pace (Alpha = .05)

! Interpret with caution--standard errors are poorly estimated.

Figure 11.2-1

AGE 9
DIFFERENCE BETWEEN BIB AND PACE PERCENTAGES

RECEIVING A SCORE GREATER THAN OR EQUAL TO 2

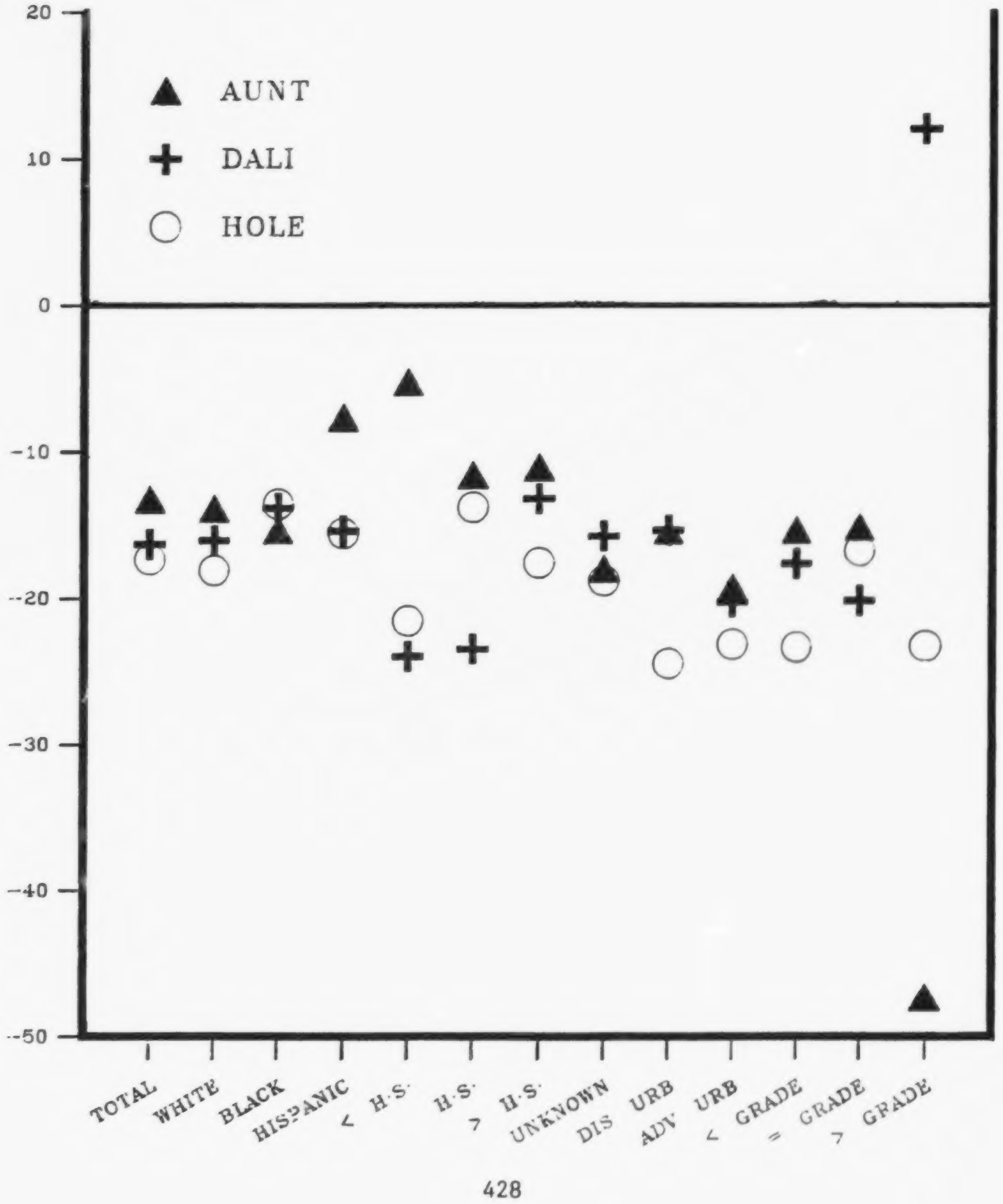


Figure 11.2-2

AGE 13
DIFFERENCE BETWEEN BIB AND PACE PERCENTAGES

RECEIVING A SCORE GREATER THAN OR EQUAL TO 2

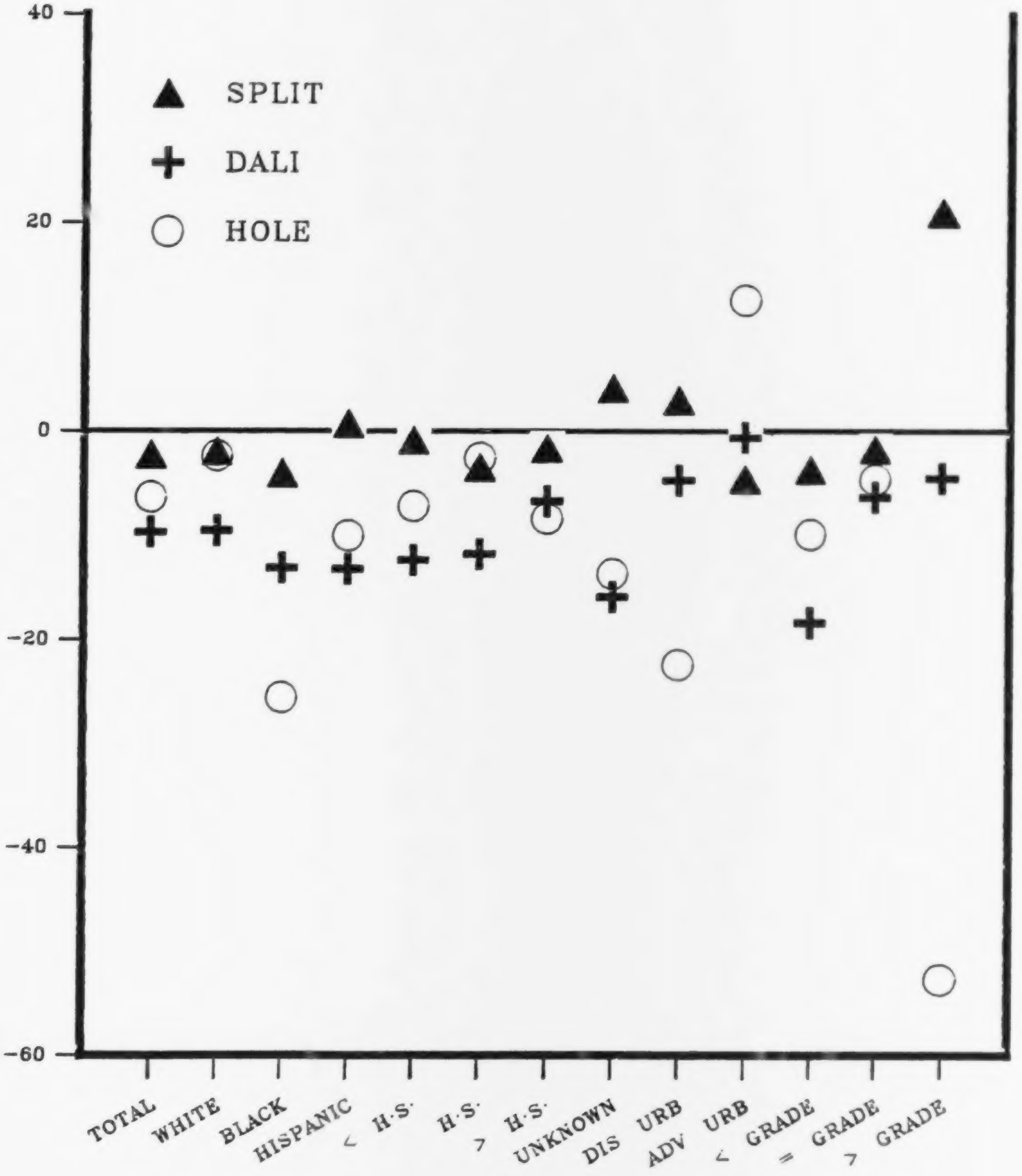
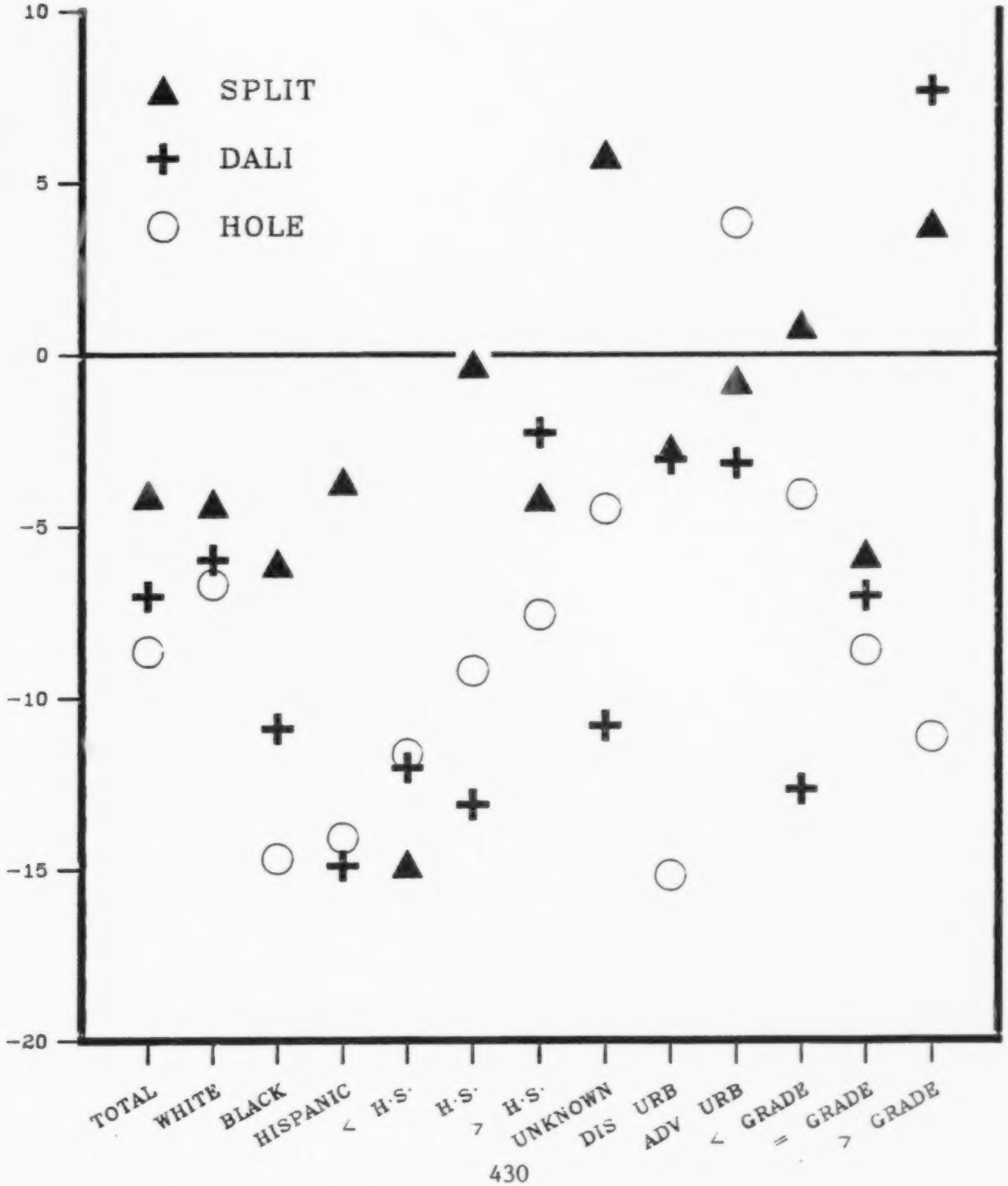


Figure 11.2-3

AGE 17
DIFFERENCE BETWEEN BIB AND PACE PERCENTAGES

RECEIVING A SCORE GREATER THAN OR EQUAL TO 2



Chapter 11.3

ESTIMATION OF TRENDS IN WRITING ACHIEVEMENT

Eugene G. Johnson

Educational Testing Service

Chapter 11.2 noted that there appears to be a differential effect of mode of administration on the estimation of writing achievement. In particular, writing exercises administered using the paced tape procedures, where the instructions are read aloud to the students, tend to be less difficult for students than the BIB spiralled administrations, where the students are required to read and understand the instructions. Furthermore, the reading of the writing assignment in a paced tape administration appears to be of more benefit for some subgroups of students than for others, where the amount of benefit depends on the item. This differential benefit makes the adjustment of scores from BIB spiralled administration to correspond to scores from a paced tape administration difficult, since a different adjustment may be required for each subgroup.

Most of the Year 15 writing assessment employed BIB spiralled administration of writing exercises, in contrast with previous assessments which used only paced tape procedures. Consequently, measurements of trends in writing achievement over time, using the results from the BIB spiralled assessment (possibly adjusted for the effect of mode of administration), will be confounded by the effects of the different mode of administration in Year 15 as opposed to the previous assessments. The degree of this confounding depends on the subgroup considered and the success of the adjustment.

To eliminate the confounded effects of mode of administration on the estimates of writing achievement, the statistics used to report trends over time are not based on the full Year 15 NAEP writing assessment, but are limited to the data obtained from the subset of writing tasks at each age that were included in the booklets administered in accordance with the paced tape procedure, in exactly the same manner as in past writing assignments.

Although the need for overlapping procedures and analyses designed to link the two methods had been anticipated by NAEP staff, only about half of the previously administered writing items (and therefore only about one fifth of all the writing items included at each age/grade level in the full Year 15 writing assessment) were selected for dual assessment, appearing in both the primary BIB spiralled assessment and in the much smaller paced tape assessment. The trend results presented in the report

Writing: Trends Across the Decade, 1974-1984 (Applebee, Langer, & Mullis, 1986a) are based upon this limited selection of writing items administered at each age, and generalizations based on the results should be viewed with caution, particularly when they pertain to one type of writing at one age level.

These writing items span different periods in NAEP's history. One of the items was included in both the Year 5 and the Year 10 writing assessments and two of these were included in the previous assessments as well as in the Year 15 assessment, thereby enabling comparisons in student performance to be made across ten years (Year 5 to Year 10 to Year 15) or across five years (Year 10 to Year 15).

To provide a fuller perspective on trends in writing proficiency during the last ten years, we have reported the newly analyzed trend information in the context of the trend data for those items collected during the earlier five-year time span (Year 5 to Year 10) and reported by the Education Commission of the States (1980). The complete set of trend results is based only on comparisons of identical writing tasks administered in the same way in at least two assessments. All responses to each task from all assessment administrations were evaluated at the same time by the same readers.

The data linking back to the first writing assessment (Year 1) were minimal: one single national sample (about 2,500 papers) on one imaginative writing task rated using the primary trait method at each age level, and one national subsample (about 400 papers) on a different task at each age level rated holistically. Given these limited data and the fact that any subgroup trends from Year 1 to Year 5 would be based on only one imaginative writing task, the writing report is limited to trends over the last decade based on changes between the Year 5, Year 10 and Year 15 assessments. The full set of writing items used is summarized in Table 11.3(1).

Table 11.3(1)
Exercises Used to Estimate Trends in Writing Performance

	Scoring Method[1]	Number in Sample								
		1974			1979			1984		
		9	13	17	9	13	17	9	13	17
INFORMATIVE										
Dali[2] (description)	P, H	-	-	-	2482	2496	2433	1351	1275	1539
Electric Blanket (business letter)	P	-	-	2276	-	-	2781	-	-	-
Describe (description)	H	-	420	417	-	536	539	-	-	-
PERSUASIVE										
Aunt May[2] (letter)	P, H	-	-	-	2525	-	-	1386	-	-
Split Session[2] (letter)	P, H	-	-	-	-	2735	2742	-	1276	1540
Puppy Letter (letter)	P	2643	-	-	2494	-	-	-	-	-
Principal (letter)	P	-	2552	-	-	2793	-	-	-	-
Recreation Center (written speech)	P	-	-	2308	-	-	2784	-	-	-
IMAGINATIVE										
Hole in the Box[2] (description)	P, H	2543	2513	2246	2464	2782	2688	1344	1289	1534
Goldfish (description)	P	2611	-	-	2475	-	-	-	-	-
Loss (description)	P	-	2607	-	-	2775	-	-	-	-
Fireflies (narration)	P	2573	-	-	2553	-	-	-	-	-
Kangaroo (narration)	H	409	-	-	494	-	-	-	-	-
Rainy Day (narration)	P	-	2621	-	-	2804	-	-	-	-
Stork (narration)	P	-	-	2281	-	-	2748	-	-	-
Grape Peeler (satire-humor)	P	-	-	2283	-	-	2765	-	-	-
BACKGROUND QUESTIONS		-	-	2237	-	29430	26631	-	5158	6209

[1] P = Primary Trait, H = Holistic

[2] Analysis performed by ETS in conjunction with analysis of the Year 15 writing assessment results

Chapter 11.4

THE AVERAGE RESPONSE METHOD (ARM) OF SCALING[1]

Albert E. Beaton
Eugene G. Johnson

Educational Testing Service

The National Assessment of Educational Progress (NAEP) used a variant of multiple matrix sampling called BIB spiralling (Beaton, 1984) in its Year 15 assessment. Multiple matrix sampling allows the assessor to administer a large number of exercises in a subject area, more exercises than would be prudent to ask any individual student to perform. BIB spiralling has the additional property of assuring that each pair of exercises is administered to a randomly equivalent subsample of students. The BIB spiralling was imposed on an already complex multi-staged sample design. In sum, the NAEP of 1983-84 contained many reading and writing exercises, as well as hundreds of questions about backgrounds, attitudes, and activities, which were collected on a sample of over 100,000 students in this nation's schools.

The results of an assessment like this would be hard to integrate and interpret if the vast array of information were presented in an exercise-by-exercise manner. NAEP has elected to summarize the available information by developing scales which encapsulate much of the information available in the exercise responses. Separate scales have been developed for the reading and writing exercises. It is the purpose of this chapter to describe the rationale and properties of the writing scale.

The properties of the reading scale have been reported in Chapter 10.5. The technology of the reading scale was not appropriate for the writing scale. For reading, there were a large number of exercises, of which 228 were used in the scale, and the individual exercises could be scored as right or wrong, so item response theoretic (IRT) methodology could be adapted for the scale. For writing, there were only 22 exercises, of which only ten were useful for the writing scale, and the individual exercises were graded on a zero to four scale, so standard IRT methodology was not appropriate. Several attempts have been made to adapt IRT technology to these non-binary writing exercise responses, but these efforts have not proved fruitful at this time.

---

[1] The statistical programming for this section was provided by Bruce Kaplan, David Freund, and Laurel Barnett. The figures were produced by Ira Sample.

Both the reading and writing portions of the assessment do have important features in common. Perhaps most important here is that the information available about most students is sparse so that the scale scores for few, if any, students are sufficiently accurate for individual decision making. A teacher or administrator would insist on a more reliable test, that is, a test with many more items in the subject area, before using the test for making decisions which would affect a student's academic career. However, a national assessment does not report individual scores and is concerned primarily with the producing national and regional parameter estimates and measures of the accuracy of estimation. The unreliability of individual scores has led Mislevy (1985a) to expand Rubin's (1978) work on missing data to assessments, and this work has been incorporated into both the reading and writing scale construction and analysis.

The writing scale is defined for NAEP as the average of a subject's scores on ten specific essays that were administered to NAEP subjects. These ten essays were chosen because they were administered at more than one age level and because all inter-correlations among them were estimable. The (unobserved) writing scale score is a latent variable, since no individual student actually responded to more than four essays and thus the average over all ten essays must be estimated. The result of the scaling process is a set of plausible values for each student who responded to at least one of the ten essays; each plausible value is a different estimate of the student's unknown writing scale score. The five different estimates for each student are values from the conditional distribution of potential scores for the student, conditional on the available information, and reflect the uncertainty in estimation.

The writing scale described here is closely related to an estimation procedure suggested by Goldstein and James (1983). Goldstein and James address the estimation of population averages of test scores where the scores are the sum of item responses, and such estimation is the primary concern of this scaling method as well. To improve the estimates, the NAEP scaling procedure uses other available information in the estimation process. The writing scale also results in the plausible values which may be considered as partial computations that can be used for estimating other parameters. Also, the partial computations are useful in estimation with a complex sample, such as NAEP's. Proper use of the plausible values allows for an accounting of the uncertainty due to incomplete information both due to the sampling of individuals from the population and due to the incomplete information on each sampled individual. However, the plausible values may result in biased estimates of parameters that were not included in the scaling process (see Sections 11.4.3 and 11.4.6).

The next two sections of this chapter will develop the scaling method. The following section will discuss the properties of the plausible values of the scale score. The final sections will discuss the specifics of the application of this technique to NAEP writing data.

11.4.1 Method

As mentioned above, the writing scale score is the average of a set of writing exercises. Let us assume that we wish to estimate the average writing score for some group, say, males. To be more general, we will assume that we wish to estimate a set of parameters called β. β may be the mean of any subgroup, a set of means, or any arbitrary parameters that may contribute to or be related to performance in writing. If we can estimate β, then we can estimate any linear combinations of β. Let us be explicit about the notation and assumptions. Let

Z	be an Nxp matrix of rank p for the writing scores of the N ($i=1,2,\ldots,N$) subjects on the p writing essays. The values z_{ik} ($k = 1, 2, \ldots, p$) will be known for those who were administered the k^{th} exercise and unknown otherwise.
a	is a p-element column vector of known constants. Although a may contain any values, we will generally assume here that all values $a_k=1/p$.
X	be an Nx(m+1) matrix containing the values of the m conditioning variables for the N subjects. The values x_{ij} ($j = 0, 1, 2, \ldots, m$) of the conditioning variables are assumed known for all subjects. The $zero^{th}$ column of X is a vector of unities. For convenience, we will use $m'=m+1$. For simplicity here, we will assume that X is of rank m'.

The Nth-order column vector Y is defined as

$Y = Za$ where the elements of Y, y_i, are the writing scale scores. The exact value of y_i will not be known unless a subject i was administered all writing tasks.

We will assume that we have identified the complete set of conditioning variables, X, and that the effect of the conditioning variables is to move the centers of the distribution while leaving the spread alone so that, to a reasonable degree of approximation,

$Z=XB+E$ where B is an m'xp matrix of unknown constants and E is an Nxp matrix of unknown errors. Also, we assume that each row of E is independently and identically distributed as $N(0, \Sigma)$.

Consequently,

$$Y=XBa+Ea=X\beta+\varepsilon$$

where β=Ba and ε=Ea. It follows that

$$\varepsilon \sim N(0,\sigma^2) \text{ where } \sigma^2 = a'\Sigma a.$$

Although some of the values of Z are unknown, the BIB spiralling procedure produces sufficient information to estimate the mean and standard deviation of each writing score and also the correlation between each pair of scores. Furthermore, because the BIB spiralling procedure presents items and pairs of items to randomly equivalent (i.e., representative--see Chapter 5) subsamples, estimates of means, variances, and covariances, based on the total set of available responses, are unbiased for the population values.

A maximum likelihood estimate of the cross-products matrix can be computed using the EM algorithm of Dempster, Laird, and Rubin (1977). After forming the matrix V=[X|Z], let

$$C = \text{ the maximum likelihood estimate of } V'V$$

where the cross-product matrix C has the expected value (E)

$$\underline{E}(C) = \underline{E}\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} = \begin{bmatrix} X'X & X'XB \\ B'X'X & B'X'XB+N\Sigma \end{bmatrix}$$

Using C, the mean and variance of the scale score y can be estimated as can its correlations with the variables in V. Consider a transformation matrix

$$T_y = \begin{bmatrix} I_{m'} & 0 & 0 \\ 0 & I_p & a \end{bmatrix}$$

where $I_{m'}$ and I_p are appropriately sized identity matrices. If V were completely known, then the N by m'+p+1 matrix

$$V_y = VT_y = [X|Z|Y]$$

would contain all of the elements of V as well as a column containing the scores y_i. Using C, the estimate of $V_y'V_y$ is

$$C_y = T_y'CT_y = \begin{bmatrix} X'X & X'Z & X'Y \\ Z'X & Z'Z & Z'Y \\ Y'X & Y'Z & Y'Y \end{bmatrix}$$

where $X'Y$, $Z'Y$, and $Y'Y$ are maximum likelihood estimates of the sums of squares and cross-products of Y with the other variables. C_y has the expected value

$$\underline{E}(C_y) = \begin{bmatrix} X'X & X'XB & X'X\beta \\ B'X'X & B'X'XB+N\Sigma & B'X'X\beta \\ \beta'X'X & \beta'X'XB & \beta'X'X\beta+N\sigma^2 \end{bmatrix}$$

The matrix C_y can be used to estimate a missing value y_i. Let z_i, the i^{th} row of Z, be partitioned $z_i = [z_{1i}|z_{2i}]$ where z_{1i} is a p_1^{th}-order vector containing the known values of z_i, z_{2i} is a p_2^{th}-order vector containing the unknown values, and $p = p_1 + p_2$. The known information of subject i can be encoded in the vector

$$v_{1i} = [x_i|z_{1i}]$$

where x_i is the i^{th} row of X. Let Z_1 be an Nxp_1 matrix of the vectors z_i, V_1 be the matrix $[X|Z_1]$, and

$$C_{1y} = \begin{bmatrix} V_1'V_1 & V_1'Y \\ Y'V_1 & Y'Y \end{bmatrix} = \begin{bmatrix} X'X & X'Z_1 & X'Y \\ Z_1'X & Z_1'Z_1 & Z_1'Y \\ Y'X & Y'Z_1 & Y'Y \end{bmatrix}$$

be the rows and columns of C_y corresponding to the columns in V_1 and Y. Then, the regression equation for estimating y from V_1 can be computed by sweeping (Beaton, 1964) the rows and columns corresponding to V_1 with the result that

$$C^{*}_{1y} = \begin{bmatrix} (V_1'V_1)^{-1} & d \\ -d' & c^{*}_{yy} \end{bmatrix}$$

where d is a $m' + p_1^{th}$ order column vector. The elements of d may also be written as two subvectors, $d'=[c_x'|c_z']$ where

$$c_x = [X'(I-K_z)X]^{-1}X'(I-K_z)Y$$

$$c_z = [Z_1'(I-K_x)Z_1]^{-1}Z_1'(I-K_x)Y$$

using $K_x = X(X'X)^{-1}X'$

and $K_z = Z_1(Z_1'Z_1)^{-1}Z_1'$.

Using d and the vector v_{1i}, it is possible to estimate the value of y_i for subject i as

$$\hat{y}_i = x_i c_x + z_{1i} c_z \quad .$$

Assuming a correctly specified model, the expected average value of $\hat{y}$ is the same as the expected average value of y but its variance is different being

$$\mathrm{var}(\hat{y}) = R^2 \sigma^2_y$$

where R^2 is the multiple correlation of y on X and Z_1 and σ_y^2 is the variance of the y_i about their mean. Thus var($\hat{y}$) is less than σ_y^2 unless $R^2 = 1$, which would indicate that the values of y were perfectly predictable from the known information. What has not been accounted for in the use of $\hat{y}$ for the prediction of y is the fact that there is a distribution of potential scores for any individual and that the estimate of $\hat{y}_i$ is, under normality assumptions, the estimated mean of the conditional distribution of the scores y_i given the known information X and Z_1. As such, $\hat{y}_i$ makes no allowance for the variability of the potential scores of an individual about the conditional mean.

This source of variability can be accounted for by estimating the variability of the residuals from the predicted values $\hat{y}_i$. (There is also variability in the prediction of $\hat{y}_i$ which will be addressed in the next section.) An estimate of the variance of the residuals about $\hat{y}_i$ is available in the term c^*_{yy} which is

$$c^*_{yy} = (Y-V_1 d)'(Y-V_1 d),$$

the residual sum of squares.

A plausible value of y, $\tilde{y}$, say, which is a realization from a distribution with the same first two moments as the distribution of y, can be formed by adding a random normal deviate $\tilde{e}$ to $\hat{y}$ where $\tilde{e}$ is normally distributed with mean 0 and variance equal to the residual mean square $C^*_{yy}/(N-m'-p_1)$. A plausible value for the respondent is then

$$\tilde{y}_i = \hat{y}_i + \tilde{e} = x_i c_x + z_{1i} c_z + \tilde{e} \; .$$

Under BIB spiralling, different subjects will be missing different scores and thus a different least squares equation will be needed for each pattern of missing data. Under the assumption that the complete set of conditioning variables has been identified and that the variability is correctly modeled, the distribution of the plausible values will, on the average, match that of y in the first two moments so that each $\tilde{y}$ value will have the same expected mean and variance as the corresponding y.

11.4.2 Using Plausible Values

Computing the plausible values provides an estimated y for each subject even though some, perhaps many, of the component parts are missing. Each value of $\tilde{y}$ is plausible, under the assumptions, and is useful in estimating the values of β or linear functions of β. However, using the values $\tilde{y}$ in least squares analyses as if they were exact values of y has some limitations. If the $\tilde{y}_i$ are used to fit a model of the form

$$\tilde{Y} = Xb + e$$

where $\tilde{Y}$ is the vector of the plausible values $\tilde{y}_i$, the matrix X is the same as defined above, b is the regression coefficient vector, and e is an error vector, the least squares estimate of

$$\tilde{\beta} = (X'X)^{-1}X'\tilde{Y}$$

is an unbiased estimate of β since

$$\underline{E}(\tilde{\beta})=(X'X)^{-1}X'\underline{E}(\tilde{Y})=\beta$$

(the proof is in the next section). Thus, $\tilde{Y}$ may be used to estimate functions of β such as group differences if group membership was coded in X and thus used in the creation of the plausible values.

Computing the usual estimate of the error in regression coefficients

$$\mathrm{Var}(\tilde{\beta}) = s_e^2(X'X)^{-1}$$

based on a single set of plausible values would result in an inaccurate accounting of the uncertainty involved in their estimation because the uncertainty in the measurement of the individual y values has not been completely accounted for. To account for this, Rubin (1978) has suggested assessing the uncertainty by repeating an analysis several times, each time using a different set of plausible values.

Given the model, there are two sources of uncertainty reflected in a set of plausible values. The first is the uncertainty in a student's score and is measured by the variability about the conditional mean score. This is addressed by the error term in $\tilde{y}_i$. The other source of uncertainty is the use of the regression equation computed from the matrix C_y. If the sample size is very large, the error introduced by using the sample regression equation as opposed to the population equation can be considered trivial and ignored. If the sampling variances and covariances are not small enough to ignore, then this uncertainty can be incorporated into the procedure by randomly selecting a value of d from a distribution of plausible d values. From least squares theory, under the above assumptions,

$$d \sim N(\gamma, \sigma_d^2 (V_1'V_1)^{-1})$$

where γ is unknown, σ_d^2 can be estimated from the data, and $V_1'V_1$ is a matrix containing known values in X and Z_1. The variance of the d can be expressed by a triangular matrix T_{v1} such that

$$T_{v1}'T_{v1} = \sigma_d^2 (V_1'V_1)^{-1}$$

Letting ε_d be a $m' + p_1$ vector of random normal (0,1) numbers, then

$$\Delta_d = \varepsilon_d T_{v1}$$

where the vector Δ_d is distributed $N(0, \sigma_d^2 (V_1'V_1)^{-1})$. To incorporate the uncertainty into the model, the vector Δ_d^d can be added to the best available estimate of γ which is $\tilde{d}$.

Rubin's recommendation, as applied here, is to generate several sets of plausible values, forming several similar data sets, and then estimating parameters using each set separately. If the uncertainty due to estimation of C_y is to be included, this uncertainty should be addressed by computing a vector Δ_d for each of the sets of plausible values and using the same vector for each plausible value within the set. Rubin shows that the average of the several sets of parameter estimates is an unbiased estimate of the parameter and that the variance of the parameter estimates is a component of uncertainty which should be added to the uncertainty due to sampling.

11.4.3 The Bias of the Average Response Method Estimator Due to Model Mis-specification

The preceding estimation technique produces plausible values $\tilde{y}$ whose first two moments match, on the average, those of the true (unobserved) value y whenever $y = X\beta + \varepsilon$ is an adequate description. This section establishes that fact and investigates the properties of the estimators $\hat{y}$ and $\tilde{y}$, computed as above, when the model is inadequate.

Suppose that, rather than the model presented in Section 11.4.2, a more adequate specification is

$$Z = XB + U\Gamma + E$$

where Z, X and B are as before, U is an N x q matrix of (potentially) known constants, Γ is a m x p matrix of unknown parameters, and E is an N x p matrix of errors, each row independently distributed as multivariate normal with means 0 and variance matrix Σ.

Then, with $Y = Z\underline{a}$ as before, the model for Y is

$$Y = X\beta + U\gamma + \varepsilon$$

where $\beta = B\underline{a}$, $\gamma = \Gamma\underline{a}$ and $\varepsilon = E\underline{a}$ is multivariate normal with zero mean and variance matrix $\sigma^2 I_n$ with $\sigma^2 = \underline{a}'\Sigma\underline{a}$.

As before, let z_i be the i^{th} row of Z, corresponding to respondent i, where z_i is partitioned as $[z_{1i}\ z_{2i}]$ with z_{1i} a 1 x p_1 vector containing the known values and z_{21} a 1 x p_2 vector containing the unknown values of z_i.

Consider the estimation of a value of y_i based only on z_{1i} and x_i, the i^{th} row of X, and ignoring the additional information in u_i, the i^{th} row of U.

Proceeding as in Section 11.4.2, form the cross-product matrix

$$C_{1y} = \begin{bmatrix} V_1' V_1 & V_1' Y \\ Y' V_1 & Y' Y \end{bmatrix}$$

where V_1 is $[X|Z_1]$. From this obtain the estimated value for respondent i as

$$\hat{y}_i = [x_i\ z_{1i}]\ (V_1'V_1)^{-1} V_1'Y.$$

For present purposes, a more convenient (but equivalent) representation of $\hat{y}_i$ is

$$\hat{y}_i = x_i (X'X)^{-1} X'Y + z_{1i.x} (Z_{1.x}'Z_{1.x})^{1} Z'_{1.x} Y_{.x}$$

where

$$Z_{1.x} = (I - X(X'X)^{-1}X')Z_1 = (I-K_x)Z_1$$

is Z_1 linearly adjusted by X,

$$Y_{.x} = (I - K_x)Y$$

is Y linearly adjusted by X, and

$$z_{1i.x} = Z_{1i} - x_i(X'X)^{-1}X'Z_1$$

is the i^{th} row of $Z_{1.x}$.

Let Γ_1 and E_1 be the columns of Γ and E corresponding to Z_1 so that

$$Z_{1.x} = U_{.x}\Gamma_1 + (I - K_x)\, E_1$$

where $U_{.x} = (I - K_x)\, U$ is orthogonal to X.

Similarly,

$$Y_{.x} = U_{.x}\gamma + (I - K_x)\, \varepsilon$$

consists of the part of the vector Y which is orthogonal to the column space of X.

Write $\beta^* = (X'X)^{-1} X'Y$ and $\alpha_1^* = (Z_{1.x}{}' Z_{1.x})^{-1} Z_{1.x}{}' Y_{.x}$

so that

$$\hat{y}_i = x_i\beta^* + z_{1i.x}\alpha_1^*$$

is the sum of two terms, each of which can be thought of as a predicted value.

The first term, $x_i \beta^*$, is the least-squares predictor of Y from X and has the expectation

$$x_i (X'X)^{-1} X'(X\beta + U\gamma) = x_i \beta + x_i (X'X)^{-1} X'U\gamma$$

where $x_i (X'X)^{-1} X'U$ is the projection of u_i on the column space of X. This prediction thus accounts for all information in Y predictable from linear combinations of the conditioning variables X but ignores any information in the part of the subspace of U which is orthogonal to X, that is, $U_{.x}\gamma$.

This information is addressed by the second predictor $z_{1i.x}\alpha^*_1$, where $z_{1.x}\alpha^*_1$ is the minimum mean squared error linear predictor of $Y_{.x}$ from from $Z_{1.x}$ alone. Since α^* is the minimum mean squared error estimator, it is in that sense optimal. Unfortunately, the estimator is also biased, as shall be seen.

Observe that for $\hat{y}_i$ to be an unbiased estimator of y_i for every i, it is required that

$$\underline{E}\,(Z_{1.x}(Z_{1.x}'\, Z_{1.x})^{-1}\, Z_{1.x}'\, Y) = U_{.x}\gamma \quad .$$

Now, since Z_1 and Y are jointly normally distributed, the conditional expectation of $Y_{.x}$, given $Z_{1.x}$, is

$$\underline{E}\,(Y_{.x} | Z_{1.x}) = U_{.x}\,(\gamma - \Gamma_1 D_{1y}) + Z_{1.x}\, D_{1y}$$

where

$$D_{1y} = \Sigma_{11}^{-1}\ \sigma_{1y},$$

$$\Sigma_{11} = \text{Var}\ (z_{1i}) \qquad \text{and}$$

$$\sigma_{1Y} = \text{Cov}\ (z_{1i}, y_i).$$

Consequently, the conditional (on $Z_{1.x}$) expectation of $Z_{1.x}\alpha^*_1$ is

$$Z_{1.x}\,(Z_{1.x}'Z_{1.x})^{-1}Z_{1.x}'\,U_{.x}(\gamma - \Gamma_1 D_{1Y}) + Z_{1.x}D_{1Y}.$$

Assuming that the sample size N is large and replacing $Z_{1.x}'\,Z_{1.x}$ and $Z_{1.x}'\,U_{.x}$ with their expectations produces

$$\underline{E}(Z_{1.x}\alpha^*_1|Z_{1.x}) \simeq Z_{1.x}\,(\Sigma_{11} + \Gamma_1'\Psi\Gamma_1)^{-1}\,\Gamma_1'\Psi\,(\gamma - \Gamma_1 D_{1Y})$$
$$+\ Z_{1.x}\,D_{1Y}\ ,$$

where

$$\Psi = (U_{.x}'\,U_{.x})/(N-m')\ .$$

Thus, unconditionally,

$$\underline{E}\,(Z_{1.x}\alpha^*_1) \simeq U_{.x}\gamma + \text{BIAS}$$

where

$$\text{BIAS} = U_{.x}\,(\Gamma_1(\Sigma_{11} + \Gamma_1'\Psi\Gamma_1)^{-1}\Gamma_1'\Psi - I)(\gamma - \Gamma_1\Sigma_{11}^{-1}\,\sigma_{1Y})$$

and so $\hat{y}_i$ is an unbiased estimator of y_i only when BIAS = 0.

The bias of $\hat{y}_i$ will be zero when any of the following three conditions obtains:

(1) $U_x \equiv 0$ so that U is contained in the column space of X (and thus all information in U is contained in X), or

(2) $\Gamma_1 \equiv 0$ and $\gamma = \Gamma\underline{a} \equiv 0$ so that the original specification $Z = XB + E$ was correct, or

(3) $\gamma - \Gamma_1\,\Sigma_{11}{}^{-1}\sigma_{1Y} = 0$ which will occur whenever $Y_{.x} = Z_{1.x}c + \delta$ where c is a vector independent of $Z_{1.x}$ and δ is a random variable independent of $Z_{1.x}$ and with expectation 0. A particular case of this is when the scores on all items are known, in which case $Y_{.x} \equiv Z_{1.x}\,\underline{a}$.

Observe that it is not sufficient to assume that $\Gamma = [\gamma\gamma\ldots\gamma]$, so that the relationships between the conditioning variables U and the vector of scores on each item is the same. For although then $\Gamma_1 = \gamma 1'_{p_1}$, the value of the bias is

$$U_{.x}\,(\gamma\gamma'\Psi f - I)\,\gamma g$$

where

$$f = 1'\,(\Sigma_{11} + (\gamma'\Psi\gamma)1\,1')^{-1}1 \qquad \text{and}$$

$$g = 1 - 1'\,\Sigma_{11}^{-1}\,\sigma_{1Y}$$

are both scalars. Under the assumption that $\Gamma_1 = \gamma 1'_{p_1}$, the bias is non-zero unless $g = 0$.

Since the predicted value $\hat{y}_i$ is biased, so is the plausible value $\tilde{y} = \hat{y} + \tilde{e}$, where $\tilde{e}$ is a random normal deviate with expectation 0 and variance $\hat{\sigma}^2_{RES}$ where

$$\hat{\sigma}^2_{RES} = \frac{1}{N-m'-p_1}\,c^*_{yy} = \frac{1}{N-m'-p_1}\,(Y_{.x}'Y_{.x} - \alpha^{*\prime}_1 Z_{1.x}'Z_{1.x}\alpha^*_1)$$

is the residual mean squared error.

Although $\hat{y}_i$ and $\tilde{y}_i$ are generally biased for each i, certain estimates which are linear combinations of the set of plausible values, one for each person, are not.

Let $BIAS_i$ be the row of the BIAS matrix corresponding to respondent i. We may write

$$BIAS_i = u_{i.x}\,\lambda_i$$

where $u_{i.x}$ is the i^{th} row of $U_{.x}$ and λ_i is the remainder of the BIAS matrix, which depends on the particular set of exercises answered by respondent i.

Then the N x 1 vector $\hat{Y}$ of predicted scores for all N respondents has expectation

$$\underline{E}(\hat{Y}) = X(X'X)^{-1}\, X'\underline{E}(Y) + U_{.x}\gamma + U_{.x}\Lambda$$

where Λ is the block diagonal matrix

$$\Lambda \quad = \quad \begin{bmatrix} \lambda_1 & 0 \ldots & 0 \\ 0 & \lambda_2 \ldots & 0 \\ 0 & 0 \ldots & \lambda_N \end{bmatrix} \quad .$$

Let L'Y be any linear combination of the unobserved scores Y which is an unbiased estimate of some population parameter (or vector or matrix of parameters) T. The corresponding estimate of T based on the predicted (or plausible) values has expectation

$$\begin{aligned} & L'\,[X(X'X)^{-1}X'(X\beta + U\gamma) + U_{.x}\gamma + U_{.x}\Lambda\,] \\ & = L'\,(X\beta + U\gamma) + L'U_{.x}\,\Lambda \\ & = T + L'U_{.x}\,\Lambda \end{aligned}$$

where $L'U_{.x}\,\Lambda$ is the bias of the estimate.

Hence, if L is a linear combination of the columns of X, so that L = XW for some W then since $L'U_{.x} = W'XU_{.x} \equiv 0$, the estimator is unbiased.

In particular, the plausible values allow the unbiased estimation of any linear function of the parameters $\beta^* = \beta + (X'X)^{-1}X'U\gamma$.

When L is not in the column space of X, the estimator $L'\tilde{y}$ provides a biased estimate of T where the amount of bias depends on the relative amount of information in $U_{.x}$ not being accounted for by the observed responses.

Measures of the potential amount of this bias for the NAEP writing plausible values are presented in Section 11.4.6. Prior to that, we turn to the application of the average response method of estimation to construct those values.

11.4.4 Application of the Average Response Method to NAEP Writing Data

The writing scale for the Year 15 assessment was based on a set of ten writing exercises and was constructed by an application of the average response method (ARM) to the observed responses to these items. ARM writing scale plausible values were computed for each student who was in one of the three modal grades (grades 4, 8 and 11) and who additionally responded to at least one of these ten writing items. This section details the construction of these plausible values.

The ten writing exercises which were selected to form the writing scale are listed in Table 11.4(1). These particular exercises were chosen because all inter-item correlations are estimable and because each of the items were administered to at least two of the grades (with one exception noted below). The selected exercises constitute the complete set of writing exercises which satisfy both of these criteria.

The letters "A" through "G" in Table 11.4(1) give the grouping of items into blocks for the purposes of administration of the items to students. These blocks are a part of the full set of nineteen blocks of items which were administered by BIB spiralling in the Year 15 assessments of reading and writing. Details of the BIB spiralling appear in Chapter 5. The pertinent characteristics for present purposes are that every block of items and every pair of blocks of items are administered to randomly equivalent subsamples of students. Approximately 2,000 students at a given grade responded to any one block of items; approximately 200 of a given grade responded to a pair of items in different blocks.

As indicated in Table 11.4(1), the entire set of ten exercises was assessed in Grade 8 while eight of the exercises were presented to students in Grade 4 and 6 to students in Grade 11. Nine of the exercises were presented to at least two grades with information on five of the exercises obtained from all three grades. (The remaining item, N000502, was presented to only Grade 8 but was included because it could be linked to item N000602 which was also given in Grade 4). The fact that not all items were presented at each grade has consequences in the estimation of the cross-product matrix C and in the estimation of plausible values. None of the students took more than four of the writing exercises and the majority took only one or two of the ten exercises. The exact distribution of the number of items taken, by grade, is shown in Table 11.4(2).

Table 11.4(1)

NAEP Writing Items for the ARM Writing Scale

Item		Block Grade 4	Block Grade 8	Block Grade 11
N000102	Dali	A	A	A
N000202	School Rule	B	B	B
N000302	Recreation Opp.	-	C	C
N000402	Food on Frontier	D	D	D
N000502	Dissecting Frogs	-	E	-
N000602	XYZ Company	E	E	-
N000702	Swimming Pool	F	F	F
N000802	Pets	F	F	-
N000902	Radio Station	G	G	-
N001002	Appleby House	G	G	G

Table 11.4(2)

Distribution by Grade of the Number of Writing Scale Items Taken by a Student

Items Taken	Number of Students Grade 4	Number of Students Grade 8	Number of Students Grade 11
1	4,570	4,261	7,979
2	2,883	3,741	2,195
3	1,022	1,966	483
4	332	1,124	0
Total	8,807	11,092	10,657

As was noted in Section 11.4.2, the basis for estimation of a predicted value for any given student is the full cross-products matrix

$$C = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}$$

from which all other necessary matrices and estimates are derived. For the construction of the NAEP writing scale, this matrix C was formed by creating an analogous matrix for each grade and then pooling the resulting matrices together.

In the matrix C, and in the grade analogues C_4, C_8 and C_{11}, the conditioning matrix X is a 0-1 design matrix specifically controlling for the main effects of the following conditioning variables:

Grade	Grade 4 Grade 8 Grade 11
Sex	Male Female
Race/Ethnicity	White Black Hispanic Other
Size and Type of Community	Advantaged urban Disadvantaged urban Other
Region	NE (Northeast) SE (Southeast) C (Central) W (West)
Parental Education	Less than High School Grad Graduated High School Post High School Unknown

The values of the conditioning variables are known for all students and so X'X in each of the cross-products matrices is directly obtained by taking the sum of squares and cross-products of the conditioning variables for each student, weighting these by the student's sampling weight and then summing across all students of the given grade.

For example, let X be the conditioning matrix for the sample of students in Grade 8. This sample was drawn using a complex sample design with unequal probabilities of selection of the various respondents in the sample. To account for these differential probabilities of selection, each student is assigned a sampling weight which is the reciprocal of the probability that the student was selected (and which also contains adjustments for nonresponse and post-stratification). An (approximately) unbiased and consistent estimate of the cross-product matrix X'X for the population of students in Grade 8 is the traditional weighted cross-product matrix

$$(X'X)_8 = \Sigma \; w_{8i} \, x_{8i}' \, x_{8i}$$

where x_{8i} is the row vector giving the values of the conditioning variables for the i^{th} student and w_{8i} is the student's weight.

Since no student responded to all writing items the remaining terms of the complete cross-product matrices C_4, C_8 and C_{11}, the terms X'Z and Z'Z cannot be directly estimated in this manner. However, the characteristics of the BIB spiralling assignment of exercises to students allows the consistent estimation of the components of these terms related to the pool of exercises assigned to a given grade. The procedure used to accomplish this estimation is discussed next. Since all ten items were presented to Grade 8, this produces the final estimate of the cross-product matrix C_8. Because not all exercises were presented to Grade 4 and Grade 11 students, additional work, discussed subsequently, is needed to estimate the matrices C_4 and C_{11}.

The cross-product matrix C_8, for Grade 8 students, is

$$C_8 = \begin{bmatrix} (X'X)_8 & (X'Z)_8 \\ (Z'X)_8 & (Z'Z)_8 \end{bmatrix}$$

The submatrix (Z'Z) is to be a consistent estimate of the 10 x 10 item score cross-product matrix that would have been obtained had all students in the Grade 8 population responded to each of the ten writing items. Typical elements of this matrix are the estimated population sum-of-squared scores for a given item (say the first) $(Z_1'Z_1)_8$ and the estimated population sum-of-products of scores for a given pair of items (say the first and second) $(Z_1'Z_2)_8$.

Because of the BIB spiral design, we can assume that the set of (approximately 2,000) students in Grade 8 who responded to a given item is a representative sample of the population (of all students in Grade 8 who would have responded to the item had it been presented to them). Consequently, the appropriately weighted sample mean, $\bar{Z}_1$, and the weighted sample variance, S_1^2, based on the total sample of students in Grade 8 responding to the first item, are consistent and unbiased estimates of the population mean and variance for that item. A consistent estimator of the sum-of-squared scores in the population of a given item (e.g., the first) is

$$(\widehat{Z_1' Z_1})_8 = W_{TOT} \; (S_1^2 + \bar{Z}_1^{\,2})$$

where W_{TOT} is the sum of weights of all Grade 8 students.

The consistent estimation of sum-of-products of scores in the population is enabled by the observation that, due to the BIB spiralling, the sample of (about 200 students in Grade 8 who responded to a given pair of items (in different blocks) is also a representative (albeit smaller) sample of the population. Consequently, the appropriately weighted sample correlation, r_{12}, based on the students in the grade who responded to both (the first and second) items is a consistent estimator of the population correlation between these items. A consistent estimate of the sum-of-products of scores on these items in the population is:

$$(\widehat{Z_1' Z_2})_8 = W_{TOT} \; (S_1 S_2 r_{12} + \bar{Z}_1 \, \bar{Z}_2)$$

where S_1 and $\bar{Z}_1$ are unbiased and based on the full set of Grade 8 students responding to item 1 and S_2 and $\bar{Z}_2$ are unbiased and based on the full set of Grade 8 students responding to item 2.

The estimation of the terms in the matrix (X′Z) was accomplished in an analogous manner.

The resultant cross-product matrix C_8 is a consistent and approximately unbiased estimator of the cross-product matrix for the population of Grade 8 students, but it is not the maximum likelihood estimator, which essentially requires estimation of the responses of each individual to the items not presented to that individual, this estimation based on the available information from that individual and the interrelationships between items observed in the entire sample. However, because the missing information (the items not presented to the individual) can be quite reasonably assumed to be missing due to a random process unrelated to the measurements of interest, the practical difference between the estimator C_8 and the maximum likelihood estimator is likely to be small and overwhelmed by the sampling variability. Actual comparison of the maximum likelihood estimator and the estimator C_8 bears this out.

Because not all items were presented at Grades 4 and 11, the cross-product matrices, C_4 and C_{11}, for those grades had missing cells, corresponding to the items which were not presented. For Grade 4, there were two missing items (N00302 and N00502). The cells of the matrix C_4 corresponding to these missing items (which includes all sums of cross-products involving either missing item) were filled in by

(1) assuming that, for the population of Grade 4 students, the conditional distribution of the two items given the background characteristics and responses to the 8 items actually assigned to Grade 4 is the same as the equivalent conditional distribution for the population of Grade 8 students, and is multivariate normal,

(2) estimating this conditional distribution from the Grade 8 sample, and

(3) combining this estimate with the known information obtained from the Grade 4 sample.

Specifically, by appropriate permutation of its rows and columns, the Grade 4 cross-products matrix C_4 can be written as

$$C_4 = \begin{bmatrix} X_4'X_4 & X_4'Z_{1(4)} & X_4'Z_{2(4)} \\ Z_{1(4)}'X_4 & Z_{1(4)}'Z_{1(4)} & Z_{1(4)}'Z_{2(4)} \\ Z_{2(4)}'X_4 & Z_{2(4)}'Z_{1(4)} & Z_{2(4)}'Z_{2(4)} \end{bmatrix}$$

where X_4 is the conditioning matrix for Grade 4 (with the dummy variables for Grade 8 and Grade 11 removed), $Z_{1(4)}$ corresponds to the set of eight items presented to Grade 4 students and $Z_{2(4)}$ to the remaining two, unpresented, items. Writing V_4 as the matrix $[X_4\ Z_{1(4)}]$ of known information for the grade, the matrix C_4 can be rewritten as

$$C_4 = \begin{bmatrix} V_4'V_4 & V_4'Z_{2(4)} \\ Z_{2(4)}'V_4 & Z_{2(4)}'Z_{2(4)} \end{bmatrix} .$$

For notational convenience, we will operate as if the entire population of Grade 4 students had been measured and that complete information by student is available for all columns of the matrix V_4. There is no loss of generality because only estimates of the terms of the cross-product matrix are required.

The cross-product matrix $V_4'V_4$ is estimated from the weighted pairwise information in exactly the same manner as was used to estimate the terms of the Grade 8 cross-product matrix C_8. However, since the items in $Z_{2(4)}$ were not presented to the Grade 4, no direct estimates of $V_4'Z_{2(4)}$ and $Z_{2(4)}'Z_{2(4)}$ are available which use only Grade 4 data. These terms are accordingly estimated on the basis of relationships present in the Grade 8 data.

Conformably permute and partition the Grade 8 cross-products matrix C_8 so that it may be written

$$C_8 = \begin{bmatrix} V_8'V_8 & V_8'Z_{2(8)} \\ Z_{2(8)}'V_8 & Z_{2(8)}'Z_{2(8)} \end{bmatrix}$$

where $V_8 = [X_8 \ Z_{2(8)}]$, X_8 being the conditioning matrix (with the Grade 4 and Grade 11 dummy variables removed and the Grade 8 dummy variable in the same position as the Grade 4 variable for the matrix X_4).

Assume that

$$[Z_{1(4)} \ Z_{2(4)}] = X_4 \ [B_{1(4)} \ B_{2(4)}] + [E_{1(4)} \ E_{2(4)}]$$

where $B_{1(4)}$ is an mx8 matrix of unknown constants,

$B_{2(4)}$ is an mx2 matrix of unknown constants,

and $[E_{1(4)} \ E_{2(4)}]$ is a matrix of unknown errors, each row of which is distributed N $(0,\Sigma)$, independent of the other rows.

Additionally, assume that

$$[Z_{1(8)} \ Z_{2(8)}] = X_8 \ [B_{1(8)} \ B_{2(8)}] + [E_{1(8)} \ E_{2(8)}]$$

where the rows of $[E_{1(8)} \ E_{2(8)}]$ are also independently N $(0,\Sigma)$ distributed.

Under these assumptions and supposing that the elements of the Grade 4 matrix V_4 were completely known, it is possible to construct an estimator of $Z_{2(4)}$ which uses the available Grade 4 data and the linear relationship between $Z_{2(4)}$ and V_8 estimated from the Grade 8 data. This estimator of $Z_{2(4)}$ is:

$$\hat{Z}_{2(4)} = V_4 \, (V_8' \, V_8)^{-1} \, V_8' \, Z_{2(8)}.$$

The corresponding estimator of the cross-product matrix $V_4' \, Z_{2(4)}$ which only requires knowledge of the matrix V_4', V_4 is:

$$V_4' \hat{Z}_{2(4)} = (V_4' V_4) \, (V_8' V_8)^{-1} \, V_8' Z_{2(8)} \, .$$

The expected value of $V_4' \hat{Z}_{2(4)}$ is:

$$\underline{E}(V_4' Z_{2(4)}) = \begin{bmatrix} \underline{E}(X_4' \hat{Z}_{2(4)}) \\ \underline{E}(Z_{1(4)}' \, \hat{Z}_{2(4)}) \end{bmatrix} = \begin{bmatrix} X_4' \, X_4 \, B^{*}_{2(4)} \\ B'_{1(4)} X'_4 \, X_4 \, B^{*}_{2(4)} + N \, \Sigma_{12} \end{bmatrix}$$

where $B^{*}_{2(4)} = B_{2(8)} + (B_{1(4)} - B_{1(8)}) \, \Sigma_{11}^{-1} \, \Sigma_{12}$

$\Sigma_{11} = \mathrm{Var}\,(E_{1(4)})$ and

$\Sigma_{12} = \mathrm{Cov}\,(E_{1(4)}, E_{2(4)})$.

This estimator is biased unless $B_{2(4)} = B^{*}_{2(4)}$.

The obvious estimator of $Z_{2(4)}' \, Z_{2(4)}$ is $\hat{Z}_{2(4)}' \, \hat{Z}_{2(4)}$ which has expected value

$$\underline{E}(\hat{Z}_{2(4)}' \, \hat{Z}_{2(4)}) = B^{*\prime}_{2(4)} \, X_4' X_4 B^{*}_{2(4)} + N \, \Sigma_{21} \, \Sigma_{11}^{-1} \, \Sigma_{12} + \underline{E}(G) \Sigma_{2.1}$$

where $\Sigma_{2.1} = \mathrm{Var}(E_{2(4)}) - \Sigma_{21} \, \Sigma_{11}^{-1} \, \Sigma_{12}$ is the conditional variance of $E_{2(4)}$ given $E_{1(4)}$ and where

$$G = \mathrm{trace}\,[(V_8' \, V_8)^{-1} \, V_4' V_4] \, .$$

Even if $B^{*}_{2(4)} = B_{2(4)}$, $\hat{Z}_{2(4)}' \, \hat{Z}_{2(4)}$ provides a biased estimate of $Z_{2(4)}' \, Z_{2(4)}$, the bias being due to reduction in variability due to prediction by regression. The value of the bias is

$$(N - \underline{E}(G))\ \Sigma_{2.1}.$$

Let $$\hat{\Sigma}_{2.1} = \frac{1}{N-m-p_1}\left(Z'_{2(8)} Z_{2(8)} - Z'_{2(8)}\ V_8 (V_8{}' V_8)^{-1} V'_8 Z_{2(8)} \right),$$

where

$m + p_1$	is the number of columns of V_8 and
$\hat{\Sigma}_{2.1}$	is the residual mean square for prediction of $Z_{2(8)}$ by X_8 and $Z_{1(8)}$, and is an unbiased estimator of $\Sigma_{2.1}$.

Further,

$$(N-G)\ \hat{\Sigma}_{2.1} \text{ is an unbiased estimator of } (N-\underline{E}(G))\ \Sigma_{2.1}\ .$$

The appropriate estimator of the cross-product matrix $Z_{2(4)}{}' Z_{2(4)}$ is

$$Z_{2(4)}{}' \hat{Z}_{2(4)} = \hat{Z}_{2(4)}{}' \hat{Z}_{2(4)} + (N - G)\ \hat{\Sigma}_{2.1}$$

which has expected value

$$\underline{E}(Z_{2(4)}{}' \hat{Z}_{2(4)}) = B^{\star}_{2(4)}{}' X_4{}' X_4 B^{\star}_{2(4)} + N \operatorname{Var}(E_{2(4)})$$

and thus is unbiased whenever $B^{\star}_{2(4)} = B_{2(4)}$.

Estimation of the missing cells in the Grade 11 cross-product matrix C_{11} was accomplished in an analogous manner, again using the relationships from the Grade 8 data.

Finally, the overall cross-product matrix C was formed by pooling the grade level cross-product matrices C_4, C_8 and C_{11}. In this pooling, the main effects of grade (i.e., the intercept term in each of the grade level cross-product matrices) were kept separate. All other conditioning effects (the main effects of race, region, size and type of community and parental education) were pooled across grades as were the item-by-item cross-product matrices.

The resultant matrix C was then used as the basis for constructing the matrix C_Y as was detailed in Section 11.4.2. The estimation of plausible values for all students in all three grades was accomplished according to the formulas in Sections 11.4.2 and 11.4.3 using the matrix C_Y as the basis. To approximately account for the effects of the sample design and the amount of information available, the matrix C_Y was scaled to be consistent with a sample size of 200. Five plausible values were computed for each student. Note that the additional source of uncertainty due to the estimation of parameters, described in Section 11.4.3, was not included. This, in effect, assumes that the sample size is large enough so that the variance contribution due to the regression parameters can be neglected.

11.4.5 An Alternate, Unbiased, Estimator for Linear Combinations of Mean Writing Scores

Section 11.4.4 showed that the ARM scale values produce estimates of composites which are generally biased with the amount of bias related to the amount of information in the neglected conditioning variables U which is not linearly contained in the employed conditioning variables X.

For a wide class of statistics an alternate estimation procedure is available which produces estimates which are unbiased, regardless of whether or not all appropriate conditioning variables have been identified.

This alternate procedure is based on the facts that

(1) the target quantity of interest, Y, is a linear combination of the component quantities $Z_1, Z_2, \ldots, Z_p$ (so that $Y = Z\underline{a}$), and

(2) information on the values of each of the item score variables, the Z_i, is available on a representative subsample of the population.

Suppose that the value of the mean score across the p items were known for every individual in the sample, so that the vector Y were completely known, and consider the statistic

$$t = L'Y,$$

for some vector or matrix L. Thus t is a linear combination of the elements in Y. Examples of this are subgroup means, contrasts of subgroup means, and more generally, regression coefficients.

Suppose that t is an unbiased estimator of the population value T. Then, since

$$T = \underline{E}(t) = \underline{E}(L'Y) = \underline{E}(L'Z)\underline{a} \ ,$$

the quantity of interest T can be expressed as a linear combination of component quantities, T_i, where T_i is the equivalent population value for the scores Z_i on item i, and where T_i is estimated unbiasedly by the statistic

$$t_i = L'Z_i .$$

(For the moment we assume that the score on item i is known for all individuals in the sample.)

As an example, if T is a vector of subgroup means of the average performance across the p items, then T_i is the vector of subgroup mean performance on the specific item i and so T is quite evidently the average of these item level mean performance vectors.

Now, although the score on item i is only known for a subsample of students, this subsample is a representative sample of the population. This means that an unbiased and consistent estimator of the item level parameter vector T_i based only on the available information from the subsample of students responding to the item is

$$t^*_i = L'_i Z^*_i$$

where Z^*_i is the vector of known scores and L'_i is the matrix of associated values, chosen so that

$$\underline{E}(t^*_i) = T_i .$$

In the example where T_i is the r x 1 vector of r subgroup mean performance levels on item i, the corresponding estimator t^*_i is the r x 1 vector of the weighted mean scores, by subgroup, across all members of the subgroup responding to the item.

Then, since t^*_i is an unbiased estimator of T_i, for each item i, it follows automatically that

$$t^* = \sum_{i=1}^{p} a_i t^*_i$$

is an unbiased estimator of

$$T = \sum_{i=1}^{p} a_i T_i .$$

It is also possible to obtain an estimate of the sampling variance of t^* by jackknifing the matrix $[t^*_1, \ldots, t^*_p]$ at the PSU level because, due to BIB spiralling, equivalent samples of the population of students within each PSU respond to each item. Let

$$\tau_k = [t^*_{1k}, \ldots, t^*_{pk}]$$

be the matrix with columns corresponding to the pseudo-replicates of the t^*_i corresponding to the k^{th} PSU pair. Then the pseudo-replicate of $t^* = [t^*_1, \ldots, t^*_p]\underline{a}$ corresponding to the k^{th} PSU pair is

$$t^*_k = \tau_k \underline{a}$$

and the jackknife variance estimate of t^*, which accounts for inter-item covariances is

$$\hat{\text{Var}}(t^*) = \sum_{k=1}^{M} (t^*_k - t^*)\,(t^*_k - t^*)'$$

which is a variance-covariance matrix of order r where r is the number of elements in the vector t^*. (For a further discussion of jackknife variance estimation see Chapter 13.)

Because the estimator t^* of group level data is computed as a linear combination of unbiased estimators of the corresponding parameters for each of the constituent part of $y = Z\underline{a}$, and because this linear combination is often a mean, the estimator t^* will be referred to as the meanparts estimator.

The meanparts estimator of some quantity of interest, say a group mean, differs from the equivalent estimator based on the ARM scale values in a fundamental way. The average response method seeks to obtain an unbiased estimate of the mean writing score for every individual (and goes further by also addressing the variability of that estimated score). If the method is successful, meaning that all potential conditioning variables have been included in the model (at least, to a practical approximation), then any statistics based linearly on these ARM plausible values are automatically unbiased. This means that the plausible values can be computed once and for all and that any subsequent analyses can be performed on these sets of plausible values, treating them as the actual values of y. (The analyses still need to be repeated for at least two sets of plausible values to correctly account for variability). This is extremely convenient for, especially, exploratory analysis.

On the other hand, the meanparts estimator never produces an estimate of an individual's scale value, but rather directly produces estimates of aggregate quantities, where it is required that those aggregates can be expressed as a linear combination of the equivalent aggregates of the constituent items. The advantage of this is that such estimates are unbiased, the disadvantage is that each separate analysis requires its own specific computation of the pertinent meanparts estimator, this computation requiring p separate computations: one for each of the items. This produces a considerable increase in the computational load required for exploratory analysis. Furthermore, the variance of the meanparts estimator can exceed that of the ARM estimator because the latter uses the available information more efficiently.

A practical compromise might be to conduct initial analysis on the ARM plausible values and use the meanparts analysis for the more critical analyses or to verify the results suggested by the ARM-based analyses.

How well this might work is addressed in the next section.

11.4.6 Comparison of the ARM-Based and Meanparts Estimators of Subgroup Mean Writing Scale Scores for the Year 15 Writing Assessment

This section compares estimates of subgroup writing performance, as measured by the subgroup mean of the ten writing scores, for estimates based on the ARM plausible values with the corresponding meanparts estimates, computed as the average of the subpopulation mean values on each of the constituent items.

For a specified subgroup G of students in a given grade, the estimate of the subgroup average writing performance, based on the ARM plausible values $\tilde{y}$, is the weighted mean of the plausible values for the subgroup, viz

$$\tilde{t}_G = \Sigma\, w_i \tilde{y}_i / \Sigma\, w_i$$

where w_i and $\tilde{y}_i$ are the sampling weight and plausible value (one of a set) for the i^{th} student of the given grade and specified subgroup and where the summations extend over all students in the grade who are also members of the subgroup. It has been noted that, unless the subgroup corresponds to a linear combination of the condition variables X, the statistic $\tilde{t}_G$ provides a biased estimate of subgroup performance.

The last section showed that if all ten exercises were presented to the grade, an unbiased estimate, the meanparts estimate, is available. Since all ten items were only presented to Grade 8 students, this is the only grade at which we have truly unbiased estimates of our defined measure of writing performance. (Provisionally unbiased estimates of relative performance of subgroups at the other grades will be discussed below).

Restricting our attention to Grade 8 for the moment, let t^*_{Gj} be the weighted mean score on item j over all students in the subgroup responding to the item, that is

$$t^*_{Gj} = \Sigma\, w^*_{ij} z_{ij} / \Sigma\, w^*_{ij}$$

where z_{ij} is the score on item i for the i^{th} student in the subgroup responding to the item, w^*_{ij} is that student's weight and where the summations extend over all students in the grade and subgroup who additionally responded to the item. Because the students in the subgroup additionally responding to the item constitute a representative subsample of all students in the subgroups, t^*_{Gj} is an unbiased estimate of the subgroup mean score on the item.

Then, since all ten items were presented to Grade 8, the unbiased meanparts estimate of subgroup performance (across the ten items) is

$$t^*_{G} = (1/10) \sum_{j=1}^{10} t^*_{Gj} \quad .$$

Analogous estimates of subgroup performance based on the ARM plausible values can also be obtained in the same manner for students in Grades 4 and 11. However, because not all ten items were presented to those grades, we cannot directly obtain the analogous meanparts estimates which pertain to the mean of the ten items. Rather than making additional assumptions about how the students would do on these unpresented items, we will instead define the meanparts estimators of writing performance at Grades 4 and 11 to be the average of the scores of the items actually presented at the grade. The meanparts estimator for Grade 4 is thus based on the mean of eight item-level statistics and the meanparts estimator for Grade 11 on the mean of six item-level statistics.

The resulting estimates of the average grade level writing performance are

	Grade 4	Grade 8	Grade 11
ARM	1.58	2.05	2.19
Meanparts	1.60	2.05	2.13

where the standard error of each of these means is .01. As they should, the ARM and meanparts estimates agree for Grade 8. The meanparts estimate for Grade 4, essentially a mean of the eight items given to that grade, is slightly larger than the ARM estimate which includes a prediction of the scores on the remaining two items. Since those two items were deemed inappropriate for Grade 4, being presumably more difficult, this difference is in the expected direction. For Grade 11, the meanparts estimate on the six items presented to the grade is lower than the ARM estimate which includes predictions for the four items given at Grade 8 but deemed inappropriate for Grade 11, being presumably too easy. The difference is again in the expected direction.

In practical terms, the bias of a subgroup performance estimate based on ARM plausible values is important only to the extent that it affects our estimate of the relative standing of the subgroup in relation to other subgroups or to the population as a whole. For simplicity, consider the estimation of the difference in performance level between the subgroup and the total population of students in the grade. We shall define the group effect to be this difference.

The ARM-based estimate of a group effect is

$$\tilde{D}_G = \tilde{t}_G - \tilde{t}_T \quad ,$$

the difference between the subgroup mean plausible value and the mean plausible value across all students in the grade.

The corresponding meanparts estimate is

$$D^*_G = t^*_G - t^*_T \quad ,$$

which is based only on the items presented to that grade.

A direct estimate of the bias of the ARM based estimate is the difference

$$\tilde{D}_G - D^*_G \quad .$$

For Grades 4 and 11, this difference also contains a component due to the estimation of missing items. This component can be assumed away by making the presumption that the group effects for the missing items on average equal the group effects for the items presented.

To compare the performance of the ARM-based group effect estimates with the equivalent meanparts group effect estimates we have computed both estimates for each of 169 subgroups, where each subgroup is defined by the response to one of 44 background and attitude questions, in the common core, which were asked of each student. The effects were computed separately for each of the three grades. Each of the background questions elicited information about the students demographic characteristics (e.g. sex, race, ethnicity, age), home environment (e.g. parental education, the presence of 25 or more books in the home), or school experience (e.g. number of book reports written).

Figures 11.4-1a, 11.4-1b and 11.4-1c are plots, for Grades 4, 8 and 11 respectively, of the ARM-based group effects versus the corresponding meanparts group effects for the 21 subgroups which correspond to the conditioning variables: sex, race/ethnicity, size and type of community, region and level of parental education. Because these variables were explicitly controlled for, the ARM and the meanparts estimates should be closely comparable. The figures show that, in general, this is the case with the relationship between the two estimates being well-described by a line which differs trivially from a line through the origin with a slope of 1. The relationship between the two estimates, while quite good, is, however, not perfect. The major discrepancies occur for the Grade 4 and Grade 11 data and, as noted on the figures, correspond to subgroups which constitute a relatively small proportion of the population. The reason that these larger discrepancies appear at Grades 4 and 11 but not at Grade 8 is because of the estimation of performance on the items not presented to Grades 4 or 11. Recall that the estimates of the four items not presented to Grade 11 was based on the relationships between those items (and the remaining six) observed at Grade 8. In essence, the ARM-based group effect for Grade 11 is a weighted average of the Grade 11 group effects for the six items presented to Grade 11 and the Grade 8 group effects for the four items not presented to Grade 11.

The variability of the plots of ARM versus meanparts group efects about the 45 degree line is due to the pooling of the cross-product matrices C_4, C_8 and C_{11} prior to estimation of plausible values. This pooling constrains the values of the group effects when averaged across grades, but does not constrain the values of the within-grade group effects. This corresponds to the assumption that the differential performance of a subgroup relative to the population is the same regardless of grade. Because of the generally low variability of the points in the plots about the 45 degree line, this assumption appears quite reasonable.

Of great interest, of course, is how the ARM-based group effects estimators perform on subgroups which were not specifically conditioned on. This is indicated in Figures 11.4-2a, 11.4-2b and 11.4-2c which show the plots, by grade, of the ARM-based group effects versus the meanparts group effect for 136 of the remaining 148 subgroups which were not explicitly conditioned on. (Each of the remaining twelve subgroups were based on fewer than 100 respondents and were removed on the grounds that effects could not be reliably estimated.)

Section 11.4.3 showed that the potential degree of the bias of the ARM-based group effect depends on the strength of the relation between the

Figure 11.4-1a

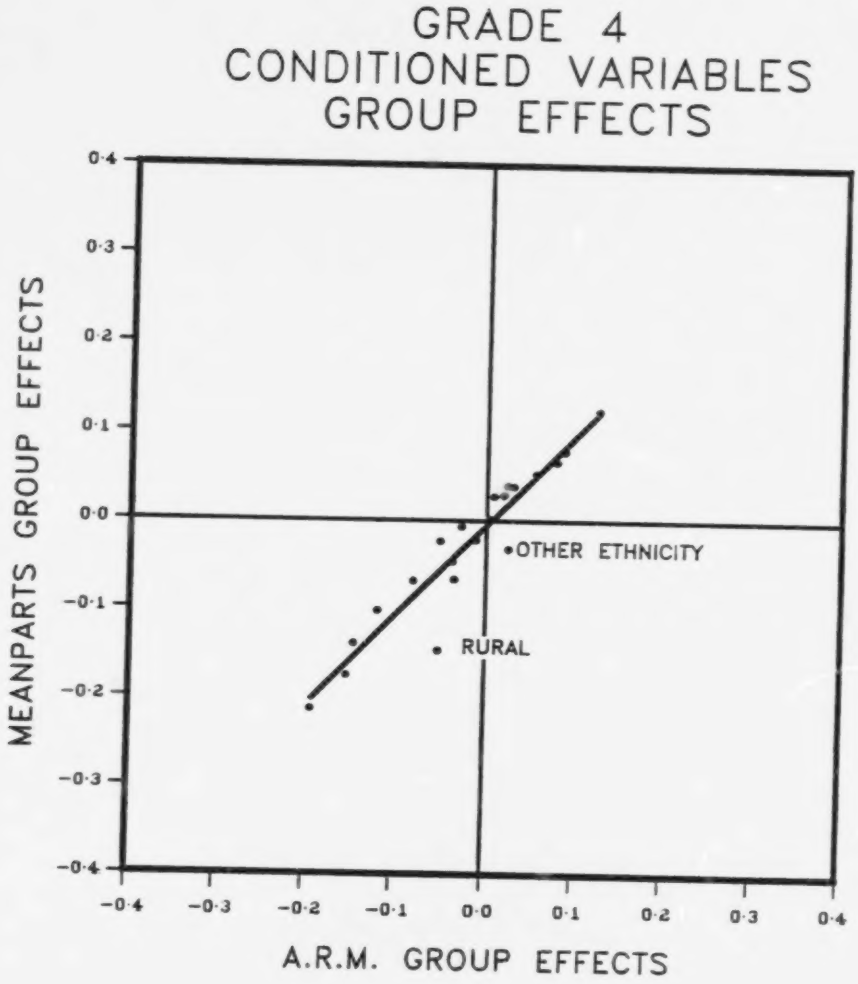


Figure 11.4-1b

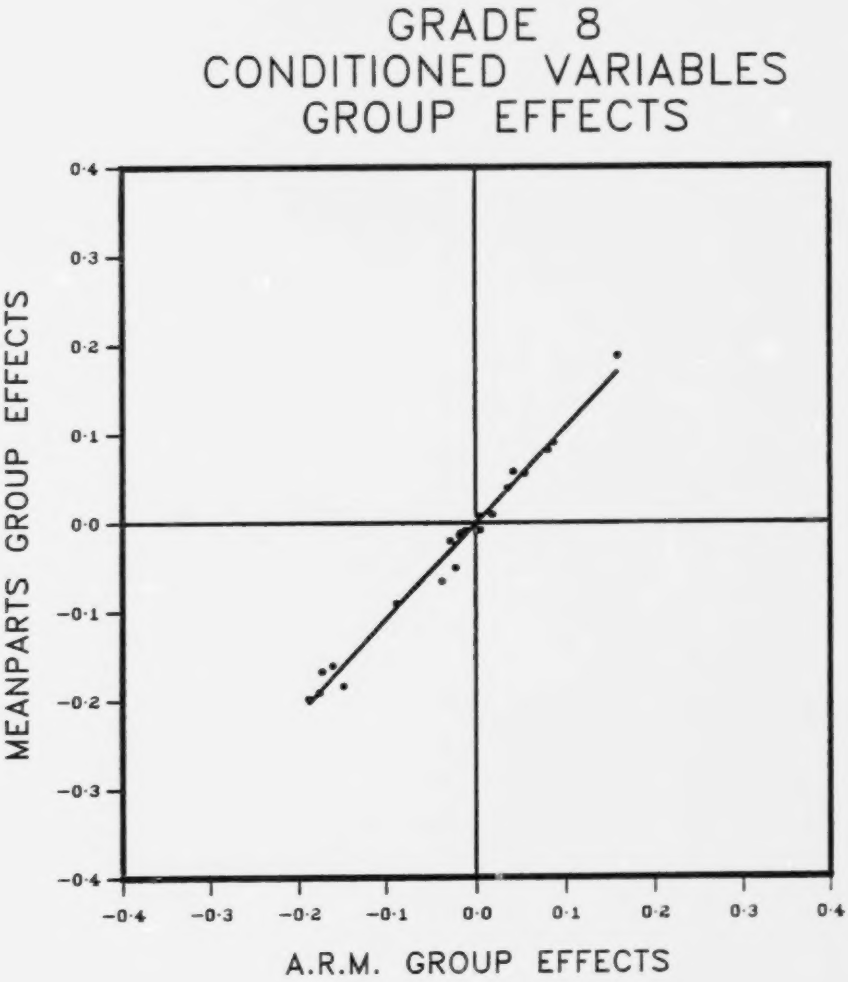


Figure 11.4-1c

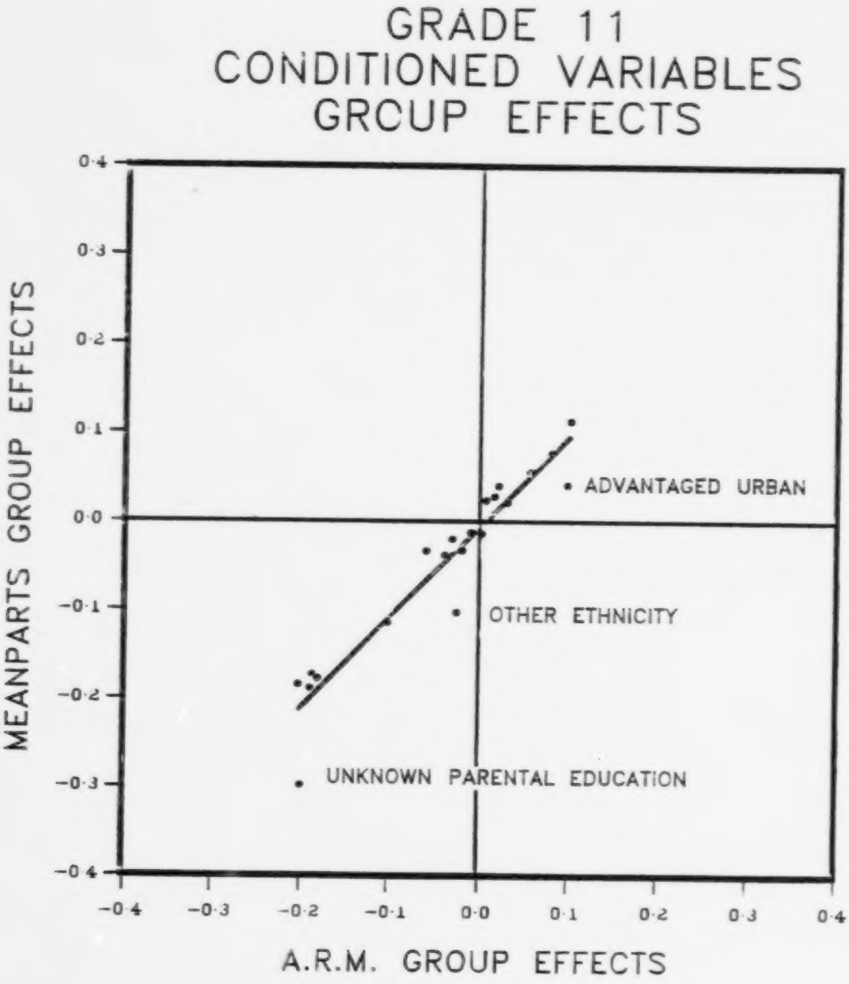


Figure 11.4-2a

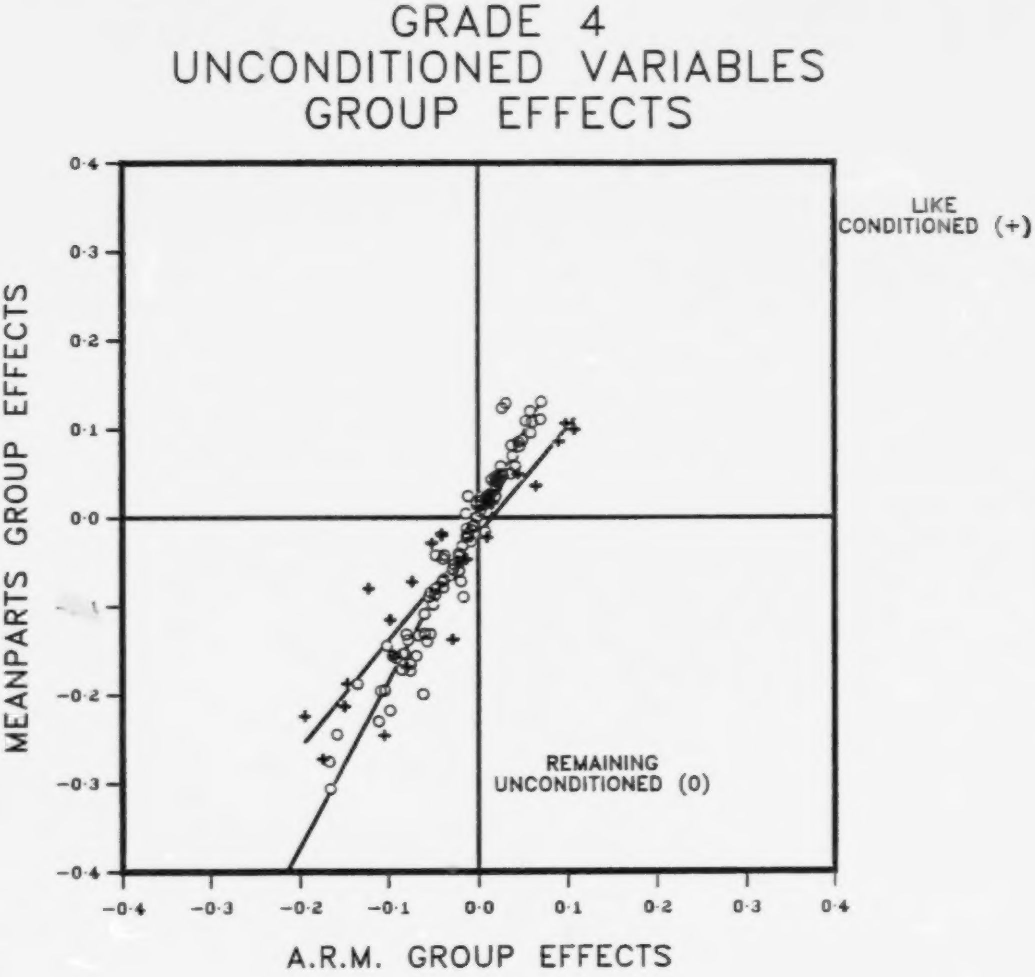


Figure 11.4-2b

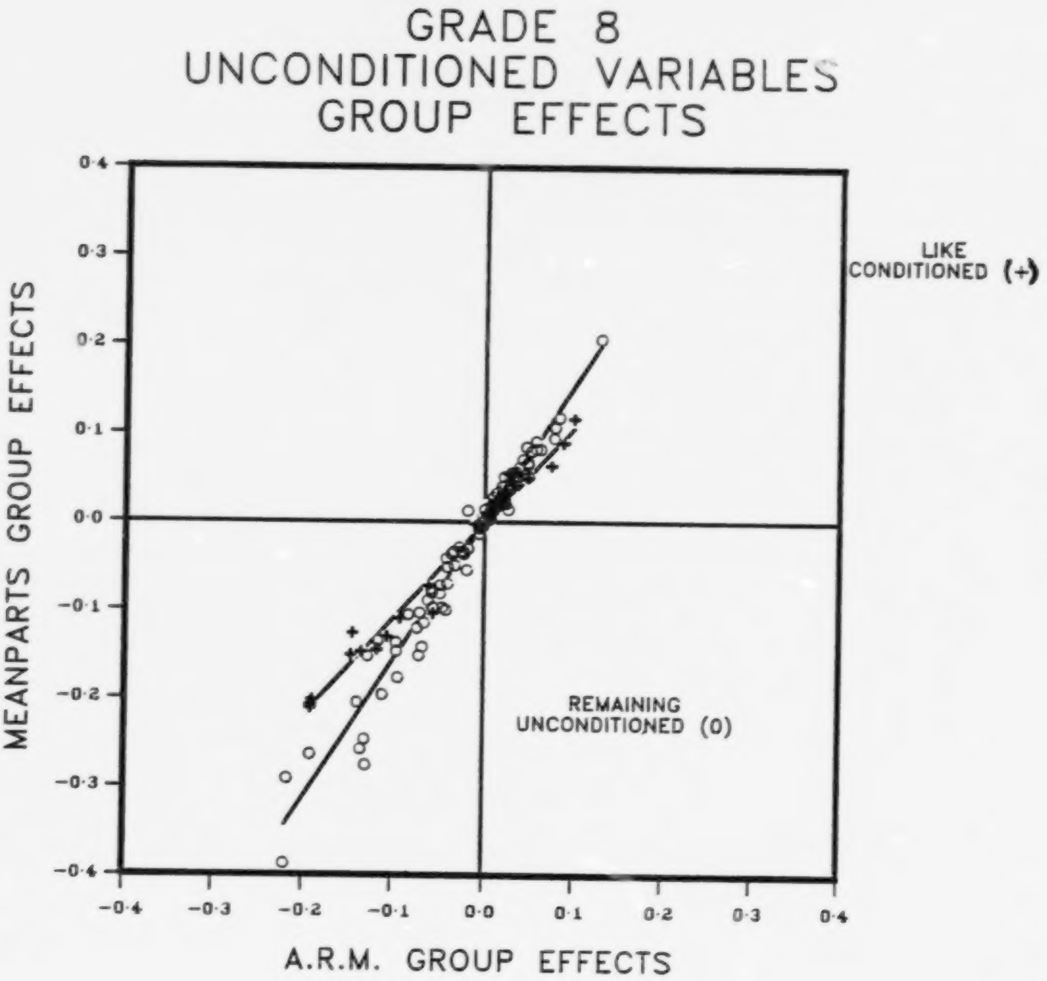
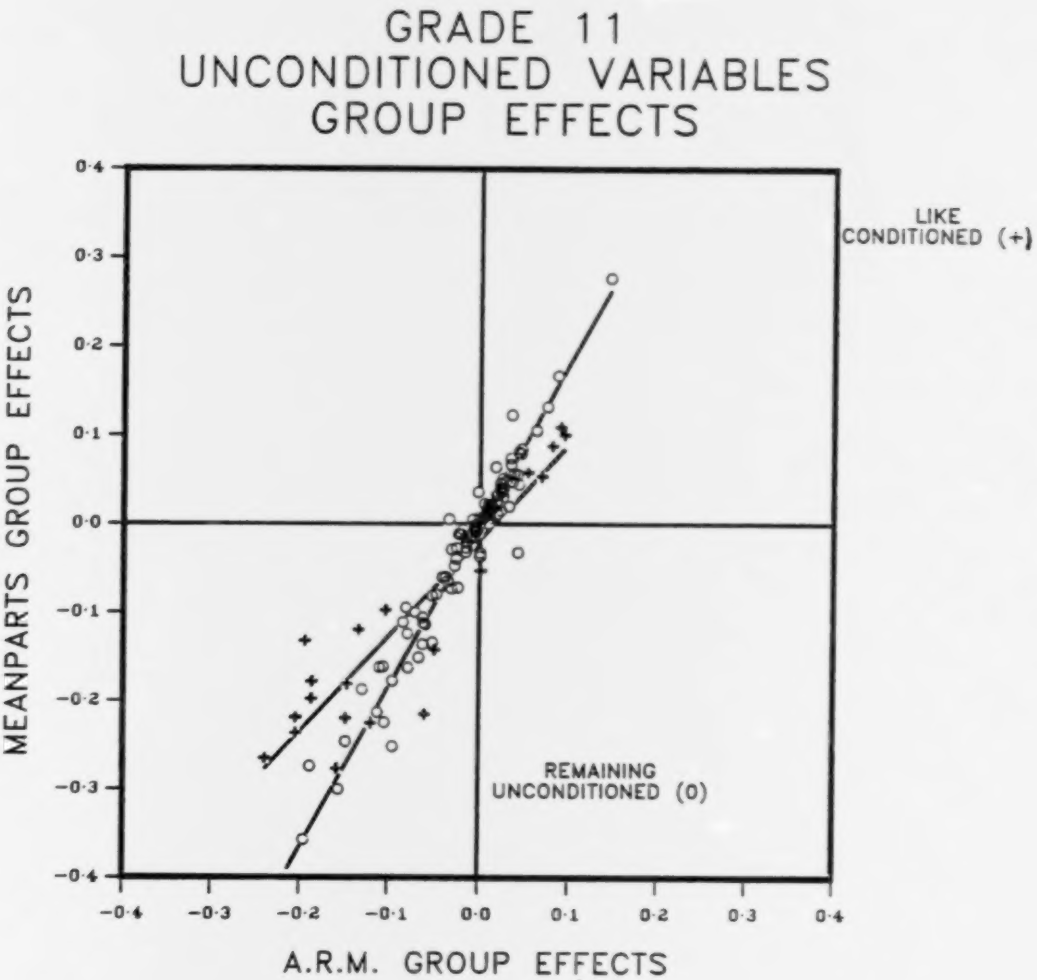


Figure 11.4-2c



group and the conditioning variables with the bias being much smaller for groups which are highly related to the conditioning variables. For this reason, we have divided the 136 subgroups addressed by Figures 11.4-2a,2b and 2c into two sets. The first set consists of the 23 subgroups formed by demographic variables which highly resemble one of the conditioning variables. Included here are subgroups based on the Level of Father's Education and on the Level of Mother's Education, both of which are used to construct the Parental Education conditioning variable. The other subgroups in the set of variables which are like the conditioning variables are related to the Race/ethnicity conditioning variable. These are: Language Spoken in the Home (English, Spanish, Other); Are You Hispanic? (No or Hispanic subgroup); and Ethnicity (American Indian/Alaska Native, Asian/Pacific Islander, Black, White, Other).

The second set of subgroups consists of the remaining 113 subgroups which are not so directly related to the conditioning variables.

Examining the plots in Figures 11.4-2a, 2b and 2c, we see that the two sets of subgroups tend to cluster along different lines. The first set, the subgroups like the conditioned variables, is indicated by +'s on the plots. The relationship between the ARM and meanparts estimates for this set tends to resemble that of the conditioned variables. This is most clear for the Grade 8 data (Figure 11.4-2b), where the "like conditioned" subgroups cluster tightly along a least-squares line with a slope of 1.09 (with a standard error of .03). The relationship between the ARM and meanparts estimates for the like conditioned subgroups for Grades 4 and 11 is not as strong as for Grade 8, but is similar. The least-squares slopes are 1.08 (standard error of .09) for Grade 11 and 1.21 (standard error .10) for Grade 4. The higher variability of the points about the lines for these two grades may be partly due to the prediction of missing information (the unpresented exercises) at those two grades.

We turn finally to the relationships between the ARM-based and meanparts group effects for the subgroups not highly related to the conditioning variables. This set of subgroups is indicated by the O's on the plots. The slopes of the least-squares lines predicting the meanparts estimate from the ARM estimate are much higher for this set of subgroups, being 1.87 for Grade 4, 1.56 for Grade 8 and 1.82 for Grade 11 (the standard errors are all about .05). This indicates a tendency for the ARM estimates to be closer to zero than the meanparts estimates so that the magnitude of subgroup effects will tend to be reduced by using the ARM estimates. The average reduction is the smallest for Grade 8 where the ARM-based estimates are around 64 percent of the magnitude of the meanparts estimates, corresponding to a shrinkage of 36 percent. The average shrinkage for the other two grades is larger, being roughly 45 percent in both cases. Again, this is due to the higher degree of prediction of missing information (the unpresented exercises) necessary for those two grades.

The general picture so far is that the estimates of group effects based on ARM plausible values will be essentially unbiased whenever the subgroups are highly related to the conditioning variables but tend to be understatements of the sizes of group effects whenever the subgroups are not highly related to the conditioning variables. We shall see that the consequences of the tendency of the ARM to understate the size of an effect, relative to the meanparts estimator, are mitigated to a large degree when the variabilities of the ARM and meanparts estimators are considered. Generally speaking, the same conclusions about subgroup effects will be made based on either of the two estimators.

Consider a test of the hypothesis of no subgroup effect for the set of subgroups defined by the responses to one of the 44 background and attitude questions. Each of these 44 questions produces a partitioning of the population into between two to ten subgroups. The analogue to the standard "F statistic" from a one-way analysis of variance, which approximately takes the sample design into account, is

$$F = \frac{1}{G-1}\left(1 - \frac{G}{n_+}\right) \frac{\sum_{i=1}^{G} f_i (t_i - t_T)^2}{\sum_{i=1}^{G} f_i (f_i - 1/n_+) \hat{\text{Var}}(t_i)}$$

where G is the number of subgroups

t_i is the (ARM or meanparts) estimate of the i^{th} subgroup mean performance

t_T is the equivalent estimate for the population (so that $t_i - t_T$ is the subgroup effect)

$\hat{\text{Var}}(t_i)$ is the estimate of the variance of t_i (which includes uncertainty in estimation of plausible values for the ARM estimate)

f_i is the weighted relative frequency of the subgroup in the population, and

n_+ is the effective sample size.

The effective sample size is the observed sample size divided by the design effect and approximately accounts for the fact that estimates of variability which take the sample design into account tend to be larger than conventional (simple random sampling based) estimates by a factor equal to the design effect (see Chapter 4 for details). For the current computations, the effective sample size was set equal to 1,000 in all cases.

To compute the approximate significance level, the above F statistic was compared with the F distribution with G - 1 and 32 degrees of freedom.

The denominator degrees of freedom, 32, is equal to the number of PSU pairs used in calculating the jackknife variance of any statistic based on the data from the Year 15 assessment and, as discussed in Chapter 4, is an upper bound for the degrees of freedom of that variance estimate. Simulation results by Shah, Holt and Folsom (1977) indicate that this is an appropriate number of error degrees of freedom to use for significance tests.

For each grade, and each of the 44 background and attitude questions, F tests in the manner described above were conducted using the ARM plausible values and using the meanparts estimates and the results compared. To eliminate the effect of different numbers of denominator degrees of freedom, the results were converted into cumulative probabilities (= 1 - significance level) and then, to facilitate plotting, into standard normal deviates. That is, the values compared were

$$Z_F = \Phi^{-1} (\text{Prob} (F_{G-1,32} \leq F))$$

where Φ is the standard normal cumulative distribution function and $\Phi^{-1} (\alpha)$ is the normal deviate at the α^{th} quantile. The results are shown in Figures 11.4-3a, 3b and 3c.

The major impression from these figures is that generally the same qualitative conclusions will be drawn from tests based on either of the ARM or meanparts estimators. As can be seen, although the ARM-based subgroup effects tend to be smaller, tests based on the ARM estimates do not appear to be markedly conservative relative to those based on the meanparts estimates, and, if anything, appear to be somewhat liberal, at least at Grade 8. The reason that the hypothesis tests, based on the ARM plausible values, are not markedly more conservative than those based on the meanparts estimates is that the estimates of the sampling variability of the ARM estimates is also smaller than the corresponding meanparts variability estimates. Figures 11.4-4a, 4b and 4c, which show the ratio of the standard errors of the ARM group performance estimate to the standard error of the meanparts group performance estimate plotted against the meanparts group performance standard error, indicate that the ARM based standard errors are, on average, around three-fourths the size of the equivalent meanparts estimates. This is true for both the conditioned and the unconditioned variables and also holds for the standard errors of group effects. The ARM-based standard errors tend to be smaller because the ARM estimators use the available information about the relationship between the exercises more efficiently than do the meanparts estimators. Specifically, the scores of an individual on the exercises not presented to that individual are partially predictable by the responses to the exercises that were answered by the individual. This means that each person provides at least some information about each of the writing exercises administered to that grade. The ARM capitalizes on this fact. The meanparts estimator does not consider this information and consequently has a larger variance.

The overall conclusion from these plots is that the general effect of the bias in the ARM-based estimates is to shrink the size of a group effect to a value of about half what it would be with the meanparts estimate but that, after taking the sampling variability into account, very few qualitative conclusions would be changed by using the ARM-based estimates rather than the much more computationally intensive meanparts estimates.

Figure 11.4-3a

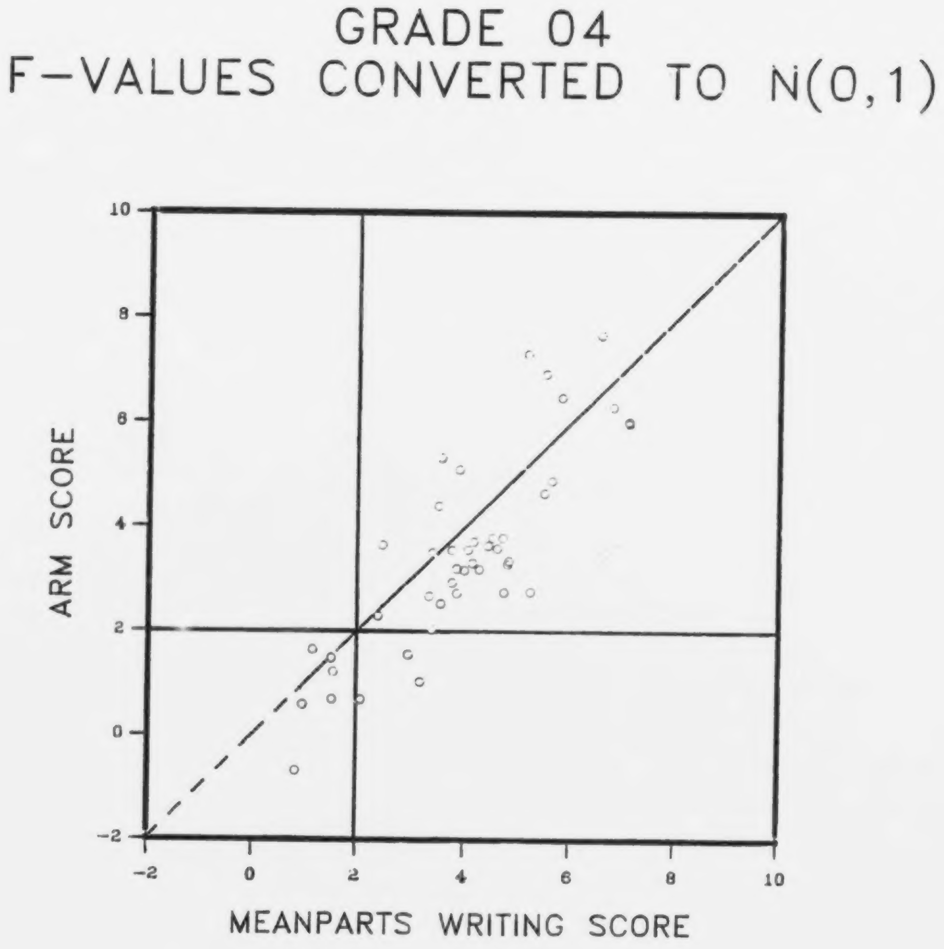


Figure 11.4-3b

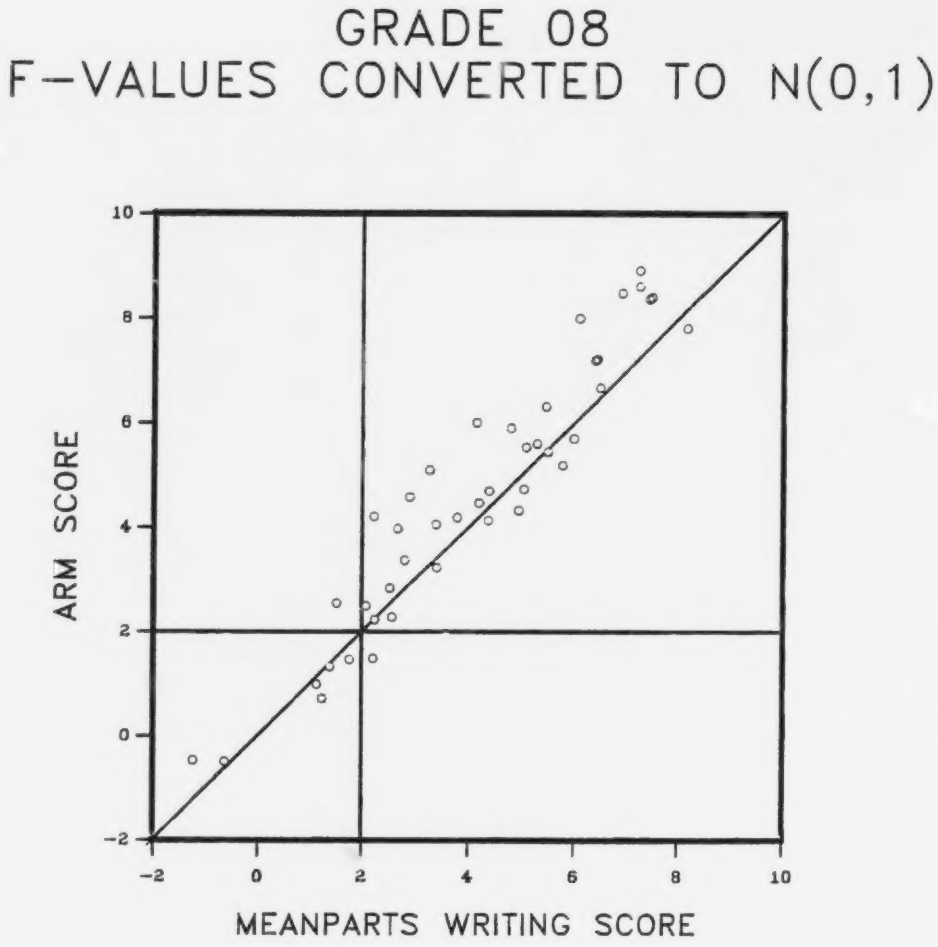


Figure 11.4-3c

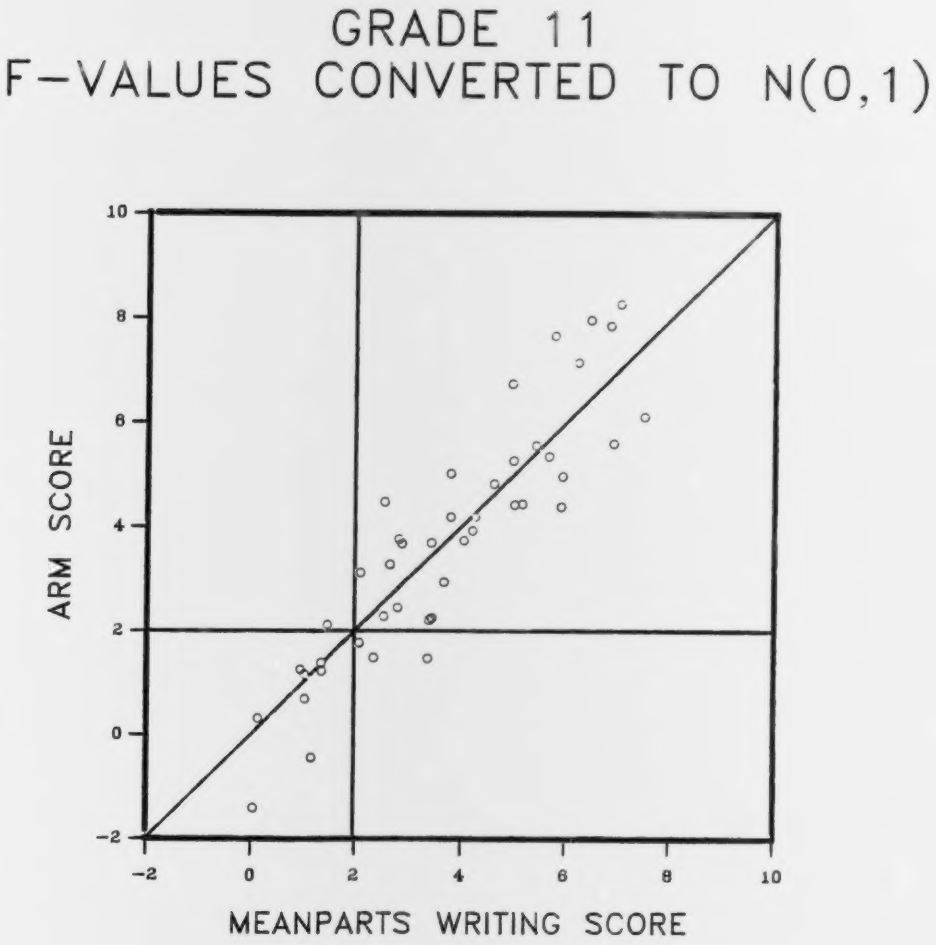


Figure 11.4-4a

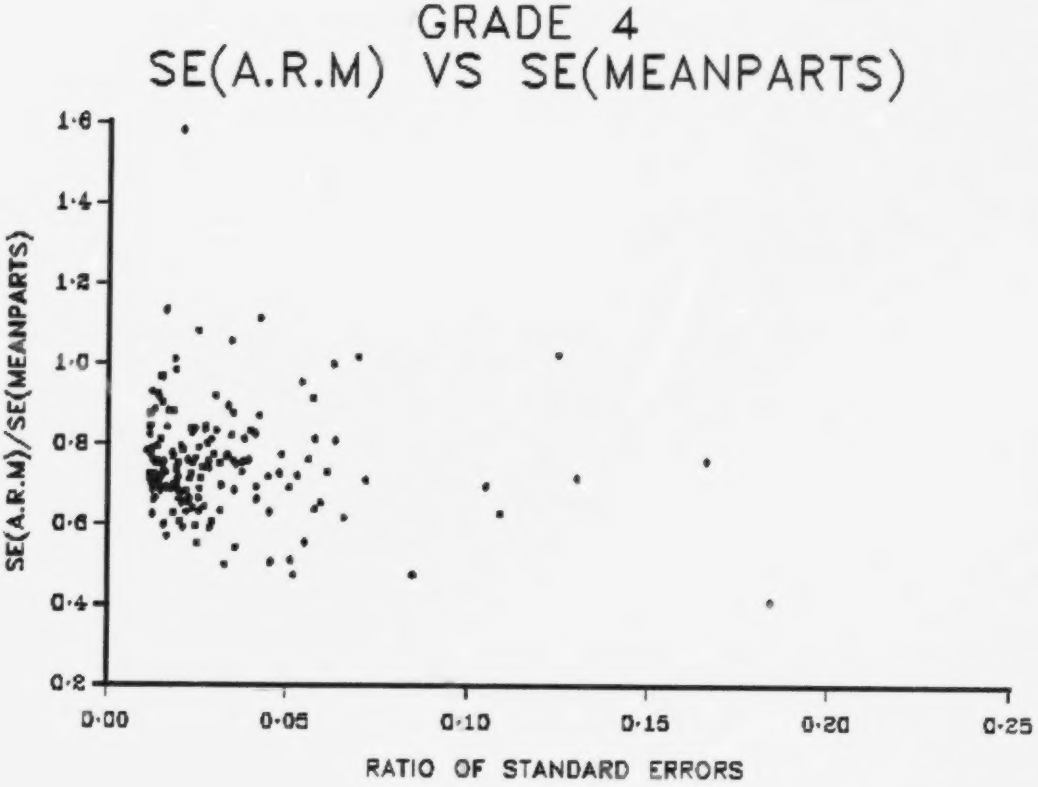


Figure 11.4-4b

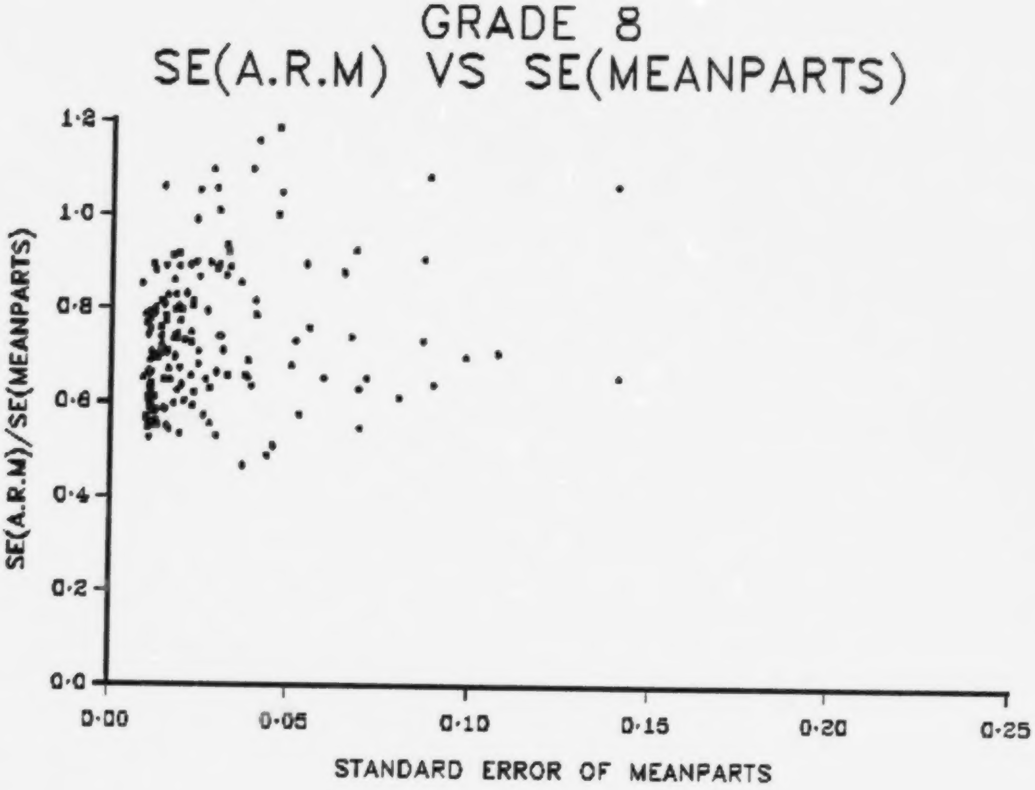
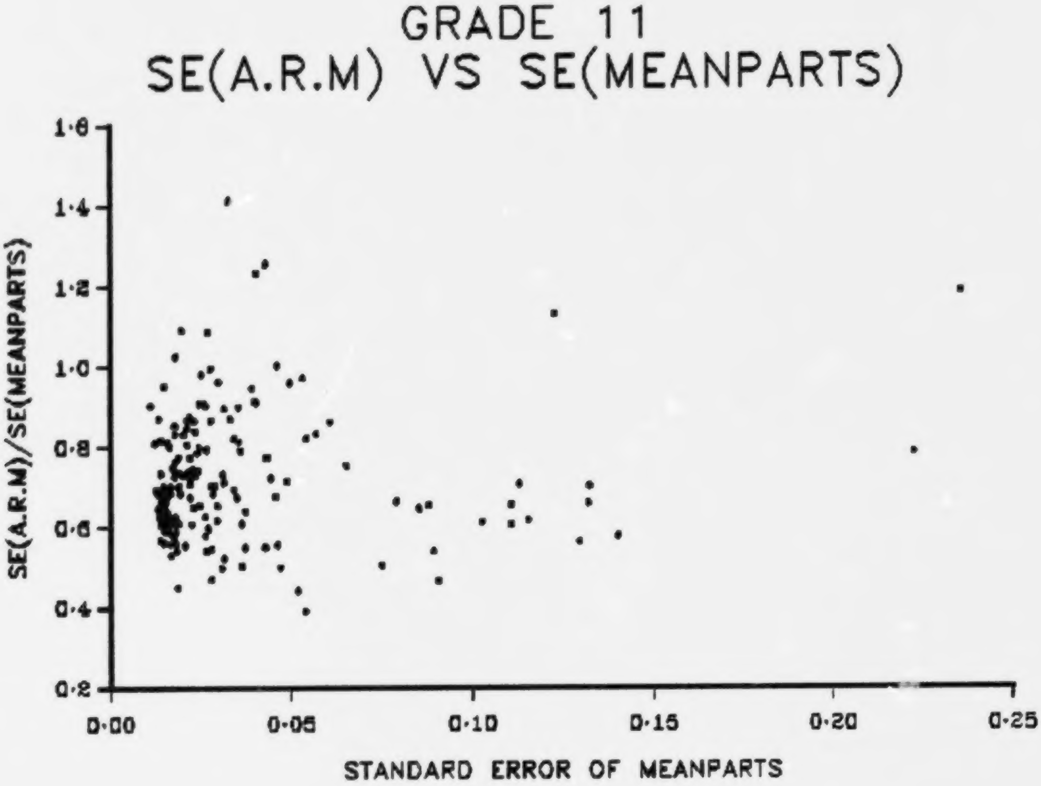


Figure 11.4-4c



Chapter 12

BACKGROUND AND ATTITUDE DATA ANALYSIS

Albert E. Beaton
Norma A. Norris
Janet R. Johnson

Educational Testing Service

The ETS design of NAEP called for the inclusion of a large number of background, attitude, and interest questions in addition to the usual cognitive exercises in the subject areas being assessed. Some questions of this type, such as the student's sex and levels of parents' education, had been asked in past assessments; these questions were continued in the Year 15 assessment. ECS supplied to ETS a large number of questions about teaching and learning styles and habits, which were also included. ETS added a large number of questions which might be useful for policy analyses. The result is a very rich database which includes not only information about reading and writing proficiency but hundreds of other variables measuring attributes of the students, their schools, and their teachers. A summary of the background and attitude questions is presented in Chapter 6.

The wealth of this database has not been fully explored by the NAEP staff, nor should it be. These data were collected for secondary analysis by persons interested in various facets of educational policy; we hope that we have developed a database sufficient for many, varied policy analyses by many researchers. The NAEP staff has devoted its energy to performing only those analyses necessary for the reports it produced.

The trend reports have been particularly limiting because they are necessarily restricted to variables that have been used in the past. Basically, these are the reporting subgroup variables of sex, race/ethnicity, region, age, grade, size and type of community, and level of parents' education. To maintain trend analysis capability, we have defined variables as closely as possible to those used in past assessments.

This chapter is divided into two sections:

- Reporting subgroups and derived variables. Section 12.1 describes the reporting subgroups and how they are defined.
- Other derived variables. The analysis of the Year 15 (1983-84) writing data incorporated a number of the questions about writing attitudes and practices. In summarizing the many

questions, several scales were developed using factor analysis and the WARM scaling method. This process is described briefly in Section 12.2.

More work on the background and attitude questions, as well as the generality of WARM scales and their properties, will continue as these questions are needed for specific analyses.

12.1 Reporting Subgroups and Derived Variables

NAEP reports performance results for groups of students rather than for individual students. In addition to reporting national results, NAEP reports information about student subgroups defined by sex, race/ethnicity (both observed and imputed), region of the country, grade/age, level of parent's education, and size and type of community.

Some subgroup data were not obtained directly from assessment responses, but were derived through procedures described in Sections 12.1.3, 12.1.6 and 12.1.7 below.

Subgroup data are contained under the variable names listed in Table 12(1).

Table 12(1)

Reporting Subgroup Variables

Subgroup	Variable Name	
	Student File	School File
Sex	SEX	-
Observed Race/Ethnicity	RACE	-
Imputed Race/Ethnicity	ETHNIC*	-
Region	REGION	SREGION
Age	STUDAGE*	-
Grade	NEWGRD	-
Size & Type of Community	STOC	SSTOC
Parent's Education	PARED*	-

* Denotes derived variable

The reporting subgroups were determined as follows:

12.1.1 Sex

Responses were reported for male and female students.

12.1.2 Observed Race/Ethnicity

This is the race/ethnicity of the student being assessed as observed by the exercise administrator. The observed definition of student race/ethnicity was the only one used in NAEP assessments prior to Year 15. This variable should be used for race/ethnicity subgroup comparisons to previous assessments.

12.1.3 Imputed Race/Ethnicity

This is an imputed definition of race/ethnicity of the student being assessed, derived from several sources of information. This variable can be used for race/ethnicity subgroup comparisons within the Year 15 assessment.

Three common background items were used to determine race/ethnicity for students who participated in the Year 15 assessment session. The items were included in every spiral assessment booklet and in each tape booklet, as follows:

Common Background Item Number 2:

2. Are you Hispanic?
    - A. No
    - B. Yes, Mexican, Mexican American, or Chicano
    - C. Yes, Puerto Rican
    - D. Yes, Cuban
    - E. Yes, Other Spanish/Hispanic (What?) ________

Students who responded to item number 2 by circling B, C, D, or E were considered Hispanic. For students who circled A, did not respond to the item, or provided information which was illegible or which could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item number 1 read as follows:

Common Background Item Number 1:

1. Are you:
    - A. American Indian or Alaskan Native
    - B. Asian or Pacific Islander
    - C. Black
    - D. White
    - E. Other (What?) ________

Students who circled A were considered American Indian; B were considered Asian; C were considered Black; and D were considered White. If a student responded by circling E, race/ethnicity was determined in accordance with the information filled in by the student as "Other (What?)."

For students who did not respond to item number 1, or who did so by providing illegible information or information which could not be classified, responses to item number 4 were examined in an effort to determine race/ethnicity. Item number 4 read as follows:

Common Background Item Number 4:
4. What language do most people in your home speak?
   A. English
   B. Spanish
   C. Another language
      (What is it?) ______

A student was considered Hispanic if he or she circled B. For a student who circled C and indicated that most people in the home spoke languages which were not English or Spanish/Hispanic, race/ethnicity was determined by classifying the language specified by the student.

For a student who did not respond to common background items 1, 2 or 4 above, observed race/ethnicity, if provided by the exercise administrator, was used.

Race/ethnicity could not be classified for a student who did not respond to background items 1, 2 or 4, and for whom an observed race/ethnicity was not provided.

The races and ethnicities which were provided by students in response to items 1, 2 and 4 above are listed in Table 12(2). Slashes indicate variations in the way races and ethnicities were spelled by students.

Table 12(3) summarizes the procedure used to determine race/ethnicity.

12.1.4 Size and Type of Community

NAEP assigned each participating school to one of seven Size and Type of Community (STOC) categories. The categories were designed to provide information about the communities in which the schools were located.

The STOC reporting categories consist of three "extreme" types of communities and four "residual" community sizes. Schools were placed into STOC categories based upon information about the type of community, the size of its population and upon an occupational profile of residents provided by school principals. The principals completed estimates of the percentage of students whose parents fit into each of six occupational categories.

Table 12(2)

Race/Ethnicity Classifications

American Indian:
- American Indian
- Cherokee
- Indianamerican/ Indiamerican
- Nativeam
- Navahoe
- Sueinda

Black:
- Afroamerican
- Black
- Blackamerican

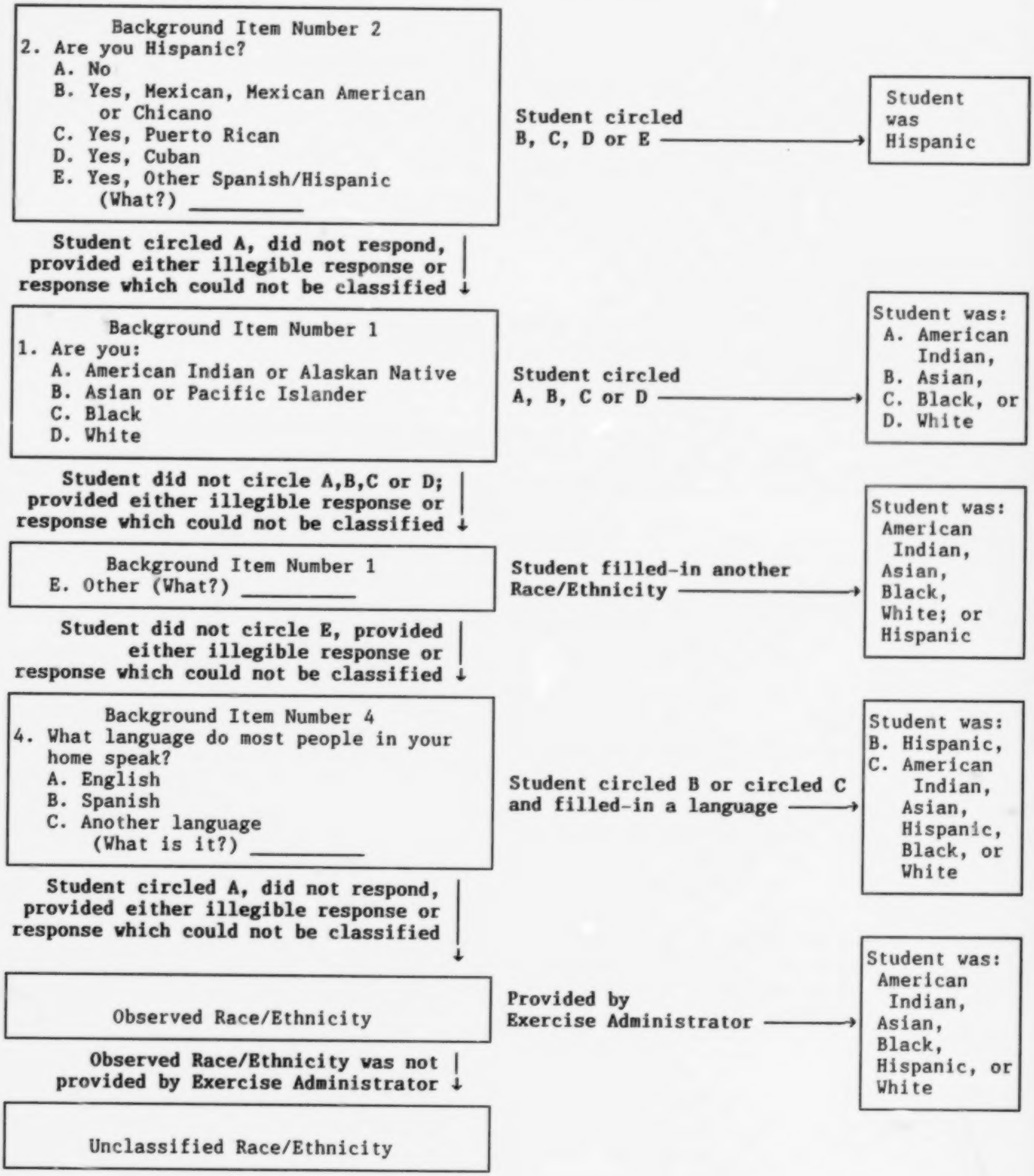
Hispanic:
- Chicano
- Columbian
- Dominican
- Hispanic
- Latin
- Mexican
- Puerto Rican
- Salvadorian
- Spanish

Asian:
- Amerasian/Amasian
- Americanphillippine
- Asian
- Assirian
- Cambodian
- Chinese
- Eastindi
- Eurasian
- Filippine/Fillippine/ Filappine/ Phillipine
- Fr India
- Guamania
- India
- Indianasian
- Indonesian
- Japanese
- Japanese-American
- Korean
- Lasos/Leocean/Laotion/ Loas
- Oriental
- Pacific
- Pakistan
- Taiwanes/Taowames
- Thai/Thia
- Vietnamese/Veitnamese

White: Unclassified:

Appropriate races/ethnicities were classified as White. Races/ethnicities which could not be considered American Indian, Asian, Black, Hispanic or White were included as unclassified.

Table 12(3)
Determining Race/Ethnicity



Background Item Number 2
2. Are you Hispanic?
   A. No
   B. Yes, Mexican, Mexican American or Chicano
   C. Yes, Puerto Rican
   D. Yes, Cuban
   E. Yes, Other Spanish/Hispanic (What?) ________

Student circled B, C, D or E →

Student was Hispanic

Student circled A, did not respond, provided either illegible response or response which could not be classified ↓

Background Item Number 1
1. Are you:
   A. American Indian or Alaskan Native
   B. Asian or Pacific Islander
   C. Black
   D. White

Student circled A, B, C or D →

Student was:
A. American Indian,
B. Asian,
C. Black, or
D. White

Student did not circle A,B,C or D; provided either illegible response or response which could not be classified ↓

Background Item Number 1
E. Other (What?) ________

Student filled-in another Race/Ethnicity →

Student was: American Indian, Asian, Black, White; or Hispanic

Student did not circle E, provided either illegible response or response which could not be classified ↓

Background Item Number 4
4. What language do most people in your home speak?
   A. English
   B. Spanish
   C. Another language (What is it?) ________

Student circled B or circled C and filled-in a language →

Student was:
B. Hispanic,
C. American Indian, Asian, Hispanic, Black, or White

Student circled A, did not respond, provided either illegible response or response which could not be classified ↓

Observed Race/Ethnicity

Provided by Exercise Administrator →

Student was: American Indian, Asian, Black, Hispanic, or White

Observed Race/Ethnicity was not provided by Exercise Administrator ↓

Unclassified Race/Ethnicity

Schools in extreme rural and low or high metropolitan areas were ranked in descending order according to the occupational profile, the type of community, and the size of its population. The top 10 percent of these schools were assigned to the extreme STOC categories (1, 2 and 3) below. The remaining schools were classified according to one of the four residual STOC categories. The three extreme STOC categories are as follows:

STOC 1 - Extreme Rural:

This category was used for schools in rural areas where a high proportion of adults were farmers or farm workers and a low proportion of professional, managerial, or factory workers. At least some of the students in these schools were from open country or places with a population of less than 10,000.

STOC 2 - Low Metro:

The low metro STOC category was used for schools in areas where a high proportion of the adult population was either not regularly employed or on welfare and a low proportion was employed in professional or managerial positions. The schools in STOC 2 were located in cities, or the urbanized area of cities, with a population greater than 200,000.

STOC 3 - High Metro:

High metro schools were located in city areas where a high proportion of adults was employed in professional or managerial positions and a low proportion factory or farm workers, not regularly employed, or on welfare. STOC 3 schools were located in cities or the urbanized area of cities with populations greater than 200,000.

Schools which did not fall into STOC 1, 2 or 3 were classified according to four "residual" STOC categories depending upon the size of the community in which they were located. The four residual STOC reporting categories are as follows:

STOC 4 - Main Big City:

STOC 4 schools were located within the limits of cities with populations greater than 200,000 but not classified as High or Low Metro.

STOC 5 - Urban Fringe:

The schools assigned to STOC 5 were located in the urbanized area, but outside the limits, of cities with populations over 200,000, but not classified as Low or High Metro.

STOC 6 - Medium City:

STOC 6 schools were located in cities with populations of between 25,000 and 200,000 which did not classify as fringe areas for big cities.

STOC 7 - Small Place:

The schools assigned to STOC 7 were located in communities with populations of less than 25,000. These communities were not located in the urbanized areas of big cities and could not be classified as Extreme Rural.

12.1.5 Region

In addition to overall responses, NAEP computed data for four geographical regions in the United States. Table 12(4) outlines the assignment of individual states to each region.

Table 12(4)
Geographic Regions

NORTHEAST:		SOUTHEAST:	
Connecticut	New Hampshire	Alabama	Mississippi
Delaware	New Jersey	Arkansas	North Carolina
District of Columbia	New York	Florida	South Carolina
Maine	Pennsylvania	Georgia	Tennessee
Maryland	Rhode Island	Kentucky	Virginia
Massachusetts	Vermont	Louisiana	West Virginia
CENTRAL:		**WEST:**	
Illinois	Missouri	Alaska	New Mexico
Indiana	Nebraska	Arizona	Oklahoma
Iowa	North Dakota	California	Oregon
Kansas	Ohio	Colorado	Texas
Michigan	South Dakota	Hawaii	Utah
Minnesota	Wisconsin	Idaho	Washington
		Montana	Wyoming
		Nevada	

12.1.6 Parental Education

Students were asked to indicate the extent of their father's education in one of the following ways:

(1) He did not finish high school;

(2) He graduated from high school;
(3) He went to another school after graduating high school;
(4) He graduated from college; or
(5) I Don't Know.

Students were asked to provide the same information about the extent of their mother's education by checking one of the following:

(1) She did not finish high school;
(2) She graduated from high school;
(3) She went to another school after graduating high school;
(4) She graduated from college; or
(5) I Don't Know.

The information was combined into one parental education reporting category, as follows:

If a student indicated the extent of education for either parent, the higher of the two levels was included in the data. If a student indicated that he or she did not know the level of education for both parents or indicated that he or she did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for either parent, the student was recorded as providing no response.

12.1.7 Grade/Age

To enhance the utility of assessment data, NAEP began sampling students by grade as well as age during the Year 15 assessment. As a result, Year 15 data reflect the following grade/age classifications:

Grade 4/Age 9 Students:

For the Grade 4/Age 9 sample, age was computed as of December 31, 1983. The sample includes many students who were both in grade 4 and age 9. However, because NAEP collected data by grade or age during the Year 15 assessment, the Grade 4/Age 9 sample also includes students who were age 9 (born in 1974) but not in grade 4, students who were in grade 4 but age 8 or younger (born in or after 1975), and students who were in grade 4 but age 10 or older (born in or before 1973).

Grade 8/Age 13 Students:

For the Grade 8/Age 13 sample, age was computed as of December 31, 1983. The sample includes many students who were both in grade 8 and age 13. However, because NAEP collected data by grade or age during the Year 15 assessment, the Grade 8/Age 13 sample also includes students who were age 13 (born in 1970) but not in grade 8, students who were in grade 8 but age 12 or younger (born in or after 1971), and students who were in grade 8 but age 14 or older (born in or before 1969).

Grade 11/Age 17 Students:

For the Grade 11/Age 17 sample, age was computed as of September 30, 1984. The sample includes many students who were both in grade 11 and age 17. However, because NAEP collected data by grade or age during the Year 15 assessment, the Grade 11/Age 17 sample also includes students who were age 17 (born between October 1, 1966 and September 30, 1967) but not in grade 11, students who were in grade 11 but age 16 or younger (born after September 30, 1967), and students who were in grade 11 but age 18 or older (born before October 1966).

12.2 Other Derived Variables

The analysis of the Year 15 writing data included not only the analysis of responses to the cognitive writing exercises, but the analysis of the responses to over 100 non-cognitive questions about the students' attitudes toward writing, their writing practices, their writing assignments, and the kind of instruction and help they received from their teachers. The many questions were reduced to a few scales using components analysis and the weighted average response method (WARM) of scaling.

First, a principal components analysis with Varimax rotation was performed to explore the dimensionality of the writing attitude and activity questions. Eleven components seemed to fit the data adequately and conform to the theory that led to these questions.

Using these components as a guide, eleven weighted sums of the items were defined as summary scales. The weighted average response method was used to estimate plausible values on each background scale for each student who had answered at least one of the items composing the scale. The weighted average response method is an extension of the average response method (ARM) which is discussed in Chapter 11.4.

A detailed description of these writing background variables is contained in the Procedural Appendix of the cross-sectional report, The Writing Report Card: Writing Achievement in American Schools, 1984 (Applebee, Langer, & Mullis, 1986b).

Chapter 13

PARAMETER ESTIMATION

Albert E. Beaton

Educational Testing Service

Given the reading, writing, and background and attitude information discussed in the preceding chapters, we can now examine what students in American schools can and cannot do. This chapter describes the process by which estimates of performance, for the nation as a whole and for selected subpopulations, were made and how the errors of estimation were produced. This chapter covers four topics:

- The weighting procedures. For any sample in which the members have different probabilities of selection, it is important to compute sampling weights for each individual. The sampling weights are used to make estimates of the parameters of the population. The NAEP sampling weights were carefully computed, using information derived from the NAEP sampling frame, from the actual NAEP data, from the Current Population Survey, and from Census Reports. The weights were developed by Westat, Inc.; the process is reported in Chapter 13.1.

- The estimation of uncertainty due to sampling variability. Each population estimate is to some degree imprecise because it is derived from a sample, and it is important to be aware of the probable magnitude of the imprecision when interpreting the results. With each parameter estimate, we have also produced an estimate of its sampling error using the jackknife method (see Mosteller & Tukey, 1969). The application of the jackknife to the NAEP data is reported in Chapter 13.2.

- The estimation of variability due to imputation. Since the NAEP sample was designed to estimate population parameters rather than individual proficiencies, individual proficiencies can not be estimated precisely. The imprecision of the individual estimates results in some additional uncertainty in the estimates of parameters. This additional uncertainty can be estimated separately, and this component of error variance can be added to the error variance due to sampling for a combined estimate of the uncertainty of parameter estimation. The steps in this process are reported in Chapter 13.3.

- The production of the basic tables of NAEP results. The basic tables consist of, among other things, estimates of the sizes of various subpopulations, the errors in estimating the sizes of subpopulations, the proportion of students able to answer each item correctly and their standard errors, and the average reading or writing proficiency of various subpopulations of students and their standard errors. Some of the tables contain trend data, that is, they compare the Year 15 data with data from past reading or writing assessments.

  These tables were designed to be informative and easy to use for the NAEP staff. They were developed over time as the staff became aware of more useful ways of presenting the data. Books of these tables, referred to as almanacs, were used as the first step in interpreting the data and serve as reference documents. The contents of the almanacs and their use are discussed in Chapter 13.4.

  The almanacs are far too voluminous to be included in this report or, indeed, to be made available to the public at large. Chapter 13.4 reports that approximately 10,000 tables have been collected into 24 books, but many more have been developed.

The statistical tables do not exhaust the parameter estimation procedures used for NAEP. Additional methods are discussed in the NAEP reports for which they were used.

Chapter 13.1

WEIGHTING PROCEDURES

Eugene G. Johnson

Educational Testing Service

Morris H. Hansen
Benjamin J. Tepping
Josefina A. Lago
John Burke

Westat, Inc.

As is the case in many large scale sample surveys, the Year 15 NAEP has a complex sample design. The goal of this design was a sample from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (as measured by low sampling variability). Additionally, it was necessary that the sample be economically and operationally feasible to obtain.

To accomplish this goal, the NAEP used a multi-stage cluster sample design (see Chapter 5) in which the probabilities of selection of the first- and second-stage sampling units (PSUs and schools) were proportional to measures of their size, but with probability for subsequent stages of sampling such that the overall probabilities of selection of students were approximately uniform, with exceptions for certain population subclasses that were oversampled by design. This oversampling was done to ensure adequate precision in the estimation of characteristics of the various subpopulations of interest. Students in the extreme rural areas and in the extreme-low-SES areas of big cities were deliberately sampled at twice the normal rate to obtain larger samples of respondents from those subpopulations. The result of these differential probabilities of selection is an achieved sample containing proportionately more members of these subgroups than there are in the population.

Appropriate estimation of population characteristics must take this disproportional representation of the various subgroups in the sample into account. This is accomplished by assigning a weight to each respondent, where the weights properly account for the sample design and reflect the appropriate proportional representation of the various types of individuals in the population.

This chapter provides an overview of the weighting procedures used for the Year 15 assessment and includes the estimation of base weights, adjustment for nonresponse, trimming of large weights, and post-stratification adjustments. Further details of these tasks can be found in the Westat Report on Sample Selection, Weighting, and Variance Estimation: NAEP--Year 15 (Lago, Burke, Tepping, & Hansen, 1985). Westat, Inc. was the subcontractor responsible for these tasks.

13.1.1 Computation of the Base Weight

The starting point for the estimation of respondent weights is the classical (Horvitz-Thompson) procedure in which the weight assigned to a respondent is the reciprocal of the overall probability that the respondent was selected for assessment. Since this weight is the basis of the final respondent weight, it is called the base weight.

The base weight assigned to a student is the reciprocal of the probability that the student was invited to a particular type of assessment session; that is, a spiral session or a particular tape session. That probability is the product of four factors:

(1) The probability that the PSU was selected;

(2) the conditional probability, given the PSU, that the school was a member of the sample selected by RTI or any supplementary sample selected by Westat;

(3) the conditional probability, given the sample of schools in a PSU, that the school was allocated the specified type of session; and

(4) the conditional probability, given the school, that the student was invited to the specified type of session.

Thus, the base weight for a student may be expressed as the product

$$W = W_1 \cdot W_2 \cdot W_3 \cdot W_4$$

where

W_1 = PSU weight,

W_2 = school weight, conditional on the PSU,

W_3 = the reciprocal of the conditional probability given the sample of schools, that the school is allocated a specified type of session, and

W_4 = the reciprocal of the within-school selection probability for students sampled for spiral or a specific tape session.

The PSU weight, W_1, was provided by RTI, the survey subcontractor of the previous grantee (ECS), and is the reciprocal of the probability of selection of the PSU. The selection probability of a given PSU was proportional to the PSU's adjusted measure of size which is ordinarily the estimated average enrollment of the three age classes. For counties identified as extreme rural, the measures of size were doubled. For big-city PSUs, the adjusted measure of size was derived as the weighted mean of the estimated enrollments for low socio-economic status tracts and for the remainder of the PSU, with the low socio-economic status tracts given twice the weight of the remainder. These adjustments were designed to effect oversampling for those counties and tracts.

The school weight, W_2, is the reciprocal of the conditional probability of selection of the school, given the selection of the PSU containing the school. This probability of selection is proportional to an adjusted measure of size for the school which is related to the estimated number of age-eligible students within the school. Roughly equal measures of size were assigned to schools containing an estimated number of age-eligible students ranging from 20 to 160 (for Age 9) or 20 to 200 (for Ages 13 and 17). Schools with fewer than 20 age eligibles were assigned smaller measures of size and schools above the indicated maximums were assigned larger measures of size which were proportional to the number of age eligibles in the school. If the school was designated as a member of the low-SES stratum of a big city, the size measure of the school was adjusted by doubling.

The session allocation weight, W_3, was computed by enumeration of all possible allocations yielded by the algorithm used by Westat to allocate tape and spiral sessions to sample schools or school clusters.

For spiral sessions, the within-school student weight, W_4, is simply the sampling interval for selecting students for spiral sessions. For tape sessions, the within-school student weight accounts for whether or not there was spiral sampling in the school and the conditional sampling interval for tape.

13.1.2 Adjustment of Base Weights for Nonresponse

The base weight for a student was adjusted by two nonresponse factors: one to adjust for non-cooperating schools and the second to adjust for students that were invited to the assessment but did not appear either in the scheduled session or in a makeup session. Thus, the within-PSU nonresponse adjusted weight was of the form

$$W_w = W_2 f_1 W_3 W_4 f_2$$

where the nonresponse adjustment factors, f_1 and f_2, were computed as described below.

The practical consequence of the nonresponse adjustments to the weights is that the distributions of characteristics of the pool of nonrespondents within a nonresponse class within a PSU are implicitly assumed to be the same, on average, as the equivalent distributions for the respondents within the same class within the PSU. That is, within classes the causes of nonresponse are in effect assumed to be ignorable so that, after appropriate adjustments of the weights, the pool of respondents can be fairly considered as a representative sample of the total population of students.

13.1.3 School Nonresponse Adjustment

A school nonresponse adjustment was applied to the base weight of students in spiral but not tape sessions, as the four required tapes per PSU were always allocated to cooperating schools in a PSU. As a result, only weights for spiral sessions were affected by school nonresponse.

School nonresponse factors were computed separately within each PSU for one, two, or three classes of schools using as many nonresponse classes as the number of sampled schools in the PSU and nonresponse pattern allowed. However, since each class was required to contain at least four or five schools, often only one class was identified in the PSU.

For any school nonresponse class, s, the school nonresponse factor for spiral sessions is given by

$$f_{1s} = \frac{\sum_{i \varepsilon A} W_{2i} G_i}{\sum_{i \varepsilon B} W_{2i} G_i}$$

where

W_{2i} = school weight (the reciprocal of the probability of selection of the school conditional on the PSU),

G_i = estimated number of grade-eligible students in school i based on QED data and/or the Principal Questionnaire,

set A consists of the original sample of eligible schools in class s (including supplemental, new, and refusing schools, but not substitutes [as defined in Chapter 4, Section 4.3]), and

set B consists of all cooperating schools in class s (including schools that were substituted for non-cooperating schools).

Note that for a substitute school, W_{2i} is the weight, based on the measure of size, which would have been used if the school had been selected by the original probability selection procedure.

13.1.4 Student Nonresponse Adjustment

Student nonresponse adjustment factors were computed separately for spiral sessions and for each of the four tape sessions within each PSU.

13.1.5 Nonresponse Adjustment for Students in Paced Tape Sessions

For each tape session, t, in a PSU, the nonresponse factor f_{2t} (≥ 1) was computed by

$$f_{2t} = \frac{n_t}{n'_t}$$

where

n_t = number of students invited to the particular tape session in the PSU, and

n'_t = number of students who completed the session.

Note that in the common situation where all students invited to a tape sessions were from a single school, no school weight (such as appears below in the adjustment factor for spiral sessions) is needed to compute the nonresponse adjustment factor; the weighted ratio equals the unweighted ratio. In the occasional situation where a school cluster was involved, it would have been appropriate to introduce the school weight in the adjustment. This was not done because of the infrequent occurrence of school clusters, and because the aggregate effect of applying the school weights in such cases would have been only marginally different from the adopted procedure of using the ratio n_t/n'_t .

13.1.6 Nonresponse Adjustment for Students in Spiral Sessions

For spiral sessions, the student nonresponse adjustment was made separately for two classes of students: those in or above the modal grade for their age and those below the modal grade.

The factor for students in class c in a particular PSU was computed by

$$f_{2c} = \frac{\sum_i W_i n_{ic}}{\sum_i W_i n'_{ic}}$$

where the summations extend over the schools in the PSU, and

n_{ic} = number of spiral invited students in school i and student class c,

n'_{ic} = number of spiral tested students in school i and student class c, and

W_i = the reciprocal of the probability of assignment of a student in school i to a spiral session, conditional on the PSU, adjusted for school nonresponse (that is, $W = W_2 f_1 W_3 W_4$).

13.1.7 Adjustment for Missing Tape Sessions

In a few instances, the supervisor inadvertently administered spiral booklets rather than the assigned tape booklet. Or, a school which was allocated a tape session refused just before the assessment was conducted, without providing enough time to reassign the tape session to another school. This problem occurred in seven of the 768 tape sessions assigned to the three grade/age groups.

The following imputation procedure was used to deal with this type of nonresponse. For variance computation purposes, the 64 NAEP PSUs had been grouped into 32 pairs. Let the PSU requiring the imputation be called the "recipient" PSU and the other member of the same pair the "donor" PSU. A one-half subsample of the students administered the particular tape session in the donor PSU was transferred to the recipient PSU. The weights of students involved in the imputation were adjusted as follows:

The students that remained in the donor PSU had their overall weight doubled by doubling the within-school student weight since this one-half subsample also represented those students transferred to the recipient PSU.

The overall weight for students in the recipient PSU was the product of its original PSU weight, the other three weights, and the student nonresponse adjustment carried from the donor PSU. The weight associated with the allocation of the particular tape session, the doubled within-school weight, and the student nonresponse adjustment were carried without modification from the donor PSU. To obtain the school weight of the recipient school, the school weight of the donor school was adjusted by the ratio of the donor PSU weight to the recipient PSU weight; that is:

$$W_{2R} = \frac{W_{1D}}{W_{1R}} \cdot W_{2D} \quad .$$

13.1.8 Trimming Extremely Large Weights to Reduce Mean Squared Error

In a number of cases, students were assigned extremely large weights. One cause of large weights was under-estimation of the number of eligible students in the school so that a school predicted to have a small number of eligible students on the basis of QED data (and hence a lower probability of selection) in fact had a large number of students. Other extremely large weights arose as the result of high levels of nonresponse coupled with low to moderate probabilities of selection.

Students with extremely large weights have an unusually large impact on estimates such as weighted means. Since the variability in weights contributes to the variance of an overall estimate by an approximate factor $1 + V^2$, where V^2 is the relvariance of the weights, a few extremely large weights are likely to produce large sampling variances of the statistics of interest, especially when the large weights are associated with students with atypical performance characteristics.

All students responding to a given type of assessment (i.e., spiral or one of the four tape assessments) within a given school receive the same weight. Consequently, extremely large weights come in groups corresponding to students in a given school. To reduce the effect of large contributions to variance from a small set of sample schools, the weights of such schools were reduced, that is, trimmed back. (We call this "weight trimming" although a more proper name would be "weight Winsorizing" to be consistent with the current terminology from robust and resistant statistics.) Following this procedure introduces a bias, but is expected to reduce the mean square error of sample estimates.

The trimming algorithm has the effect, approximately, of trimming the weight of any school that contributes more than a specified proportion, θ, to the estimated variance of the estimated number of students in the population. The trimming was done separately for the spiral assessment and for each of the four tape assessments. Let

M = number of schools in which the assessment was done,

W'_i = weight assigned to school i (i.e., the product of the PSU weight, the school weight, the school session weight and the school nonresponse factor--$W_1 W_2 W_3 f_1$),

x'_i = estimated number of age-eligible students in school i (i.e., the sum of the within-school weights for the students assessed--$W_4 f_2$ summed across the students assessed),

$x''_i = W'_i x'_i$ = number of students in the population represented by the school, and

$\bar{x}'' = (1/M) \sum_{i=1}^{M} x''_i$ = (1/M)(estimated total number of age eligibles in the population).

A rough approximation to the variance of $\bar{x}''$ is

$$\frac{1}{M} \sum_i (x''_i - \bar{x}'')^2$$

Westat adopted a trimming method that reduced the weight W'_i for a small number of schools in such a manner that no school makes a contribution to the sum shown above that is greater than a specified proportion θ, where θ is to be determined. That is, for any school j, the weight W'_j, after all weights have been trimmed if required, satisfies the condition

$$(x''_j - \bar{x}'')^2 \leq \theta V$$

where

$$V = \sum_i (x''_i - \bar{x}'')^2 \text{ is the between school sum of squares.}$$

Because only large weights are to be trimmed, the weight is not to be altered if $x''_j < \bar{x}''$.

Equivalently, the condition on the school weight W'_j is

$$W'_j \leq \frac{1}{x'_j} \left[\bar{x}'' + \sqrt{\theta V} \right]$$

The trimming was done iteratively. Using the initial weights, the weight for each school which failed to satisfy the inequality was reduced to the value given by the right-hand side of the inequality. Using the weights as trimmed, the procedure was iterated.

To determine a value of θ, the schools in each sample were listed in descending sequence according to the value

$$\theta_j = (x''_j - \bar{x}'')^2 / V \quad .$$

For alternative values of θ, it was determined how many schools violated the inequality and what their characteristics were in terms of the prior estimate of the number of eligible students, the number found to be eligible, and possibly other factors. The value of θ to be used was then chosen by judgment to provide negligible bias while substantially reducing variance. The chosen value of θ was 10/M, which resulted in a trimming of the weights for schools as follows:

	Number of schools				
Grade/Age	Spiral Assessment	Tape Assessments, by Booklet 64	65	66	67
4/9	11	0	0	0	1
8/13	3	0	0	1	0
11/17	2	1	1	2	0

Since the number of schools assigned a spiral assessment was 580 for Grade 4/Age 9, 453 for Grade 8/Age 13 and 312 for Grade 11/Age 17, the percents of schools whose weights were trimmed were 1.9 percent, .6 percent, and .6 percent, respectively. The corresponding numbers of schools assigned at least one tape session were 251 for Grade 4/Age 9, 205 for Grade 8/Age 13 and 205 for Grade 11/Age 17 (a school could be assigned one or more spiral and one or more tape sessions). The percents of schools whose weights for the paced administrations (combined) were trimmed were .4 percent, .5 percent, and 2 percent, respectively.

13.1.9 Post-stratification

As in most sample surveys with cluster sampling, the sums of respondent weights are random variables which are subject to sampling variability. Even if there were no nonresponse, the sums of the respondent weights would at best provide unbiased estimates of various subgroup totals. However, since unbiasness refers to average performance over the possible replications of the sampling, it is unlikely that any given estimate, based on the achieved sample, will exactly equal the population value. Furthermore, the respondent weights have been adjusted for nonresponse and a number of extreme weights have been reduced.

To reduce the mean squared error of the sample estimates, these weights were further adjusted so that estimated population totals for a number of specified subgroups of the population, based on the sum of weights of students of the specified subgroup, were the same as presumably better estimates derived from other sources. The details of this adjustment, which is called post-stratification, appear below.

Post-stratification replaced the "weight smoothing" that was done in the prior NAEP assessments and has the purpose (as did weight smoothing) of reducing the mean squared error of the estimated averages or proportions relating to student subpopulations that span several subgroups of the whole population. The post-stratification was done separately for the spiral sessions and each of the four tape sessions within each grade/age group, because each of these can be viewed as separate samples of the appropriate population.

For the spiral assessment, thirteen subgroups were defined in terms of race, ethnicity, census region, and community size (SDOC) as shown in Table 13.1(1). Each of the thirteen subgroups was further divided into three classes:

(a) students eligible by both age and grade;
(b) students eligible by age only;
(c) students eligible by grade only.

This resulted in 39 post-stratification cells for each age class. The final weight for a student is the product of the base weight (as adjusted for nonresponse and after "trimming") and a post-stratification factor whose denominator is the sum of those weights for the cell to which the student belongs and whose numerator is an adjusted estimate, based on more reliable data, of the total number of students in the cell.

The adjusted estimate of the total number of students in a given cell is a composite of estimates from the Year 15 NAEP sample and independent estimates based on projections based on Current Population Survey data and 1980 Census data. The adjusted estimate is a weighted mean of the two estimates, the weights being inversely proportional to the approximate variances of the NAEP and independent estimates. (Further details are provided in the Report on Sample Selection.)

The sample of students in each of the tape assessments was much smaller than the sample for the spiral assessments. Consequently, some subgroups were collapsed for post-stratification as follows:

1, 2
3
4
5
6, 7
8, 9
10, 11, 12
13

Furthermore, there was no subdivision into eligibility classes, so that there were eight post-stratification cells for each age class. The numerators of the post-stratification factors for these cells were the corresponding adjusted estimates used for computing the spiral post-stratification factor. For each of the four tape assessments, the denominators were the sums of the weights for each age class.

Table 13.1(1)

Major Subgroups for Post-Stratification

Subgroup	Race	Ethnicity	Region	SDOC*
1	White	Non-Hispanic	NE	1, 2
2	White	Non-Hispanic	NE	3, 4, 5
3	White	Non-Hispanic	SE, Central	1, 2
4	White	Non-Hispanic	SE, Central	3
5	White	Non-Hispanic	SE, Central	4, 5
6	White	Non-Hispanic	West	1, 2
7	White	Non-Hispanic	West	3, 4, 5
8	Any	Hispanic	NE, SE, Central	Any
9	Any	Hispanic	West	Any
10	Black	Non-Hispanic	NE	Any
11	Black	Non-Hispanic	SE	Any
12	Black	Non-Hispanic	Central, West	Any
13	Other	Non-Hispanic	Any	Any

*SDOC (Sample Description of Community) categories: 1--Big City; 2--Fringe of Big City; 3--Medium City; 4--Small Place; and 5--Extreme Rural.

13.1.10 The Final Student Weight: The Full-Sample Weight

The final weight assigned to a student is the student full-sample weight. This weight is the student's base weight after the application of the various adjustments described above in Sections 13.1.2 through 13.1.9.

The student full-sample weight was used to derive all estimates of population and subpopulation characteristics which have been presented in the various NAEP reports, including simple estimates such as the proportion of students of a specified type who would respond in a certain way to an exercise and more complex estimates such as mean proficiency levels.

The estimation of the variability of these estimates, however, involves the use of another set of weights, in fact, 32 other weights in all. These weights are closely related to the student full-sample weight, but differ in a manner which greatly facilitates the estimation of sampling variability by the jackknife variance estimation technique. These weights and the jackknife estimator are discussed in the next chapter.

Chapter 13.2

ESTIMATION OF UNCERTAINTY DUE TO SAMPLING VARIABILITY

Eugene G. Johnson

Educational Testing Service

A major source of uncertainty in the estimation of the value in the population of a variable of interest exists because information about the variable is obtained only on a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic. (The estimation of variability due to imperfect measurement, discussed in Chapter 13.3, is also essential).

Estimates of sampling variability are designed to provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another, equivalent, sample of individuals drawn in exactly the same manner as the achieved sample. Consequently, the estimation of the sampling variability of any statistic must take into account the design of the sample.

The NAEP sample is obtained via a stratified multi-stage probability sampling design which includes provisions for sampling certain subpopulations at higher rates. Additional characteristics of the sample include adjustments of the weights for both nonresponse and post-stratification. The resulting sample has very different statistical characteristics from those of a simple random sample. In particular, because of the effects of cluster selections (PSUs and schools within PSUs) and because of effects of nonresponse and post-stratification adjustments, observations made on different students cannot be assumed to be independent of each other (and are, in fact, generally positively correlated).

Treatment of the data as a simple random sample, with disregard for the special characteristics of the NAEP sample design, will tend to produce underestimates of the true sampling variability.

13.2.1 Linear and Nonlinear Estimators

The statistics which are obtainable from a sample can be grouped into two major types: linear and nonlinear. A linear statistic can always be represented as a sum of the form $\Sigma a_i X_i$ where the X_i are linear combinations of the observations and the a_i are fixed constants. A nonlinear statistic is anything else.

For definiteness in what follows, let $t(\underline{y}, \underline{w})$ be any statistic which is a function of the sample responses $\underline{y}$ and the weights $\underline{w}$ (both vectors). The statistic t provides an estimate of some population value of interest T. Because of the adjustments for nonresponse and the adjustments from post-stratification, the adjusted weights are random variables and consequently aggregate estimators of the $\Sigma W_i Y_i$ are nonlinear estimators. Moreover, even if the weights were not adjusted, estimates of ratios of the form $\Sigma W_i Y_i / \Sigma W_i X_i$ are nonlinear, as are the more complex estimators based on item response theory. The nonlinearity of these estimators complicates the evaluation of their sampling variability.

The sampling variability of the nonlinear estimates from the NAEP data is estimated by a jackknife procedure. The particular jackknife methodology used will be detailed below. For an explanation of the concept of the jackknife see Mosteller and Tukey (1969).

A property of jackknife methodology is that, when properly applied to the same data, a jackknife estimate of the variability of a linear estimator will produce the same result as the standard textbook variance estimate. Additionally, the jackknife estimator is a continuous function of a nonlinear estimator. Because of these properties, approximate characteristics of the jackknife estimator in the nonlinear situation (to a first-order degree of approximation) can be inferred from the characteristics in the linear situation.

13.2.2 Accounting for the Effects of Clustering, Stratification and Systematic Selection

Because the NAEP respondents are obtained by multi-stage cluster sampling, the variance of any estimate t is composed of components of variability due to each of the stages of selection. Furthermore, this variance should account for the fact that the selection of the units within PSUs at each stage is by systematic sampling and (except for the last stage) with probabilities proportional to measures of size. Appropriate estimation of the sampling variability of an estimate is aided by the remarkable and convenient fact that variance estimates based on the differences between PSU estimates also appropriately reflect the variability within PSUs, no matter how the subsampling was done, as long as the subsample taken within each PSU is a probability sample and does not depend on the subsample taken in another PSU. (For a discussion see Wolter, 1985, Section 2.4.5.; Hansen, Hurwitz, & Madow, 1953, p. 258.)

Estimation of the sampling variability of a statistic t thus comes down to the estimation of the variance between PSUs (within strata) of the sample estimates for these PSUs. Appropriate estimation by this approach will reflect approximately the combined effect of the between- and within-PSU contributions to variance. The sample of PSUs was obtained by sampling with inclusion probability proportional to an adjusted measure of size without replacement, using the algorithm developed by Chromy (1979). Since the selection was based on geographically ordered lists of PSUs within 19 sampling strata based on region by size and type of community, this produced a sample with a reasonable geographic representation. For the purposes of variance estimation, we have followed the common practice of pairing the PSUs in a manner consistent with the sample design and then regarding each pair as the members of a pseudo-stratum for variance estimation purposes. This results in a set of PSU pairs where the PSUs within a pair are nearly always both from the same stratum and tend to be geographically close to each other. Since there are 64 PSUs in total, this results in 32 pairs.

13.2.3 Estimation of Variability of Any Statistic by the Jackknife

We now turn to the general procedure used by ETS to estimate the sampling variability of any statistic t ($\underline{y}$, $\underline{w}$) which is a function of sample values $\underline{y}$ and weights $\underline{w}$. As noted above, this is done by a jackknife procedure.

As was commented in the last section, for the estimation of the sampling variability, it is sufficient to restrict one's attention to the estimation of variability of the sample estimates for each pair of PSUs in the sample. The jackknife method estimates the sampling variability of any statistic as the sum of components of variability which may be attributed to each of the PSU pairs. The variance attributed to a particular PSU pair is measured by estimating how much the value of the statistic would change if the information embodied in the PSU pair were to be changed.

This is done by the computation of a quantity t_i called a pseudo-replicate, which is associated with the i^{th} PSU pair, and which is an estimate of the statistic of interest t based on an altered sample. Specifically, the i^{th} pseudo-replicate of the statistic t is created by eliminating the data from the first PSU of the pair, replacing the lost information with that from the second PSU of the pair (so that the second PSU is included twice), and then re-estimating the statistic based on this altered set of data.

The jackknife estimate of the variability of the statistic t used by ETS is the sum of the squared differences between each pseudo-replicate and the overall value:

$$\hat{\text{Var}}\,(t) = \sum_{i=1}^{M} (t_i - t)^2$$

where M = 32 is the number of PSU pairs.

In practical terms, the major expenditure of resources in the computation of a jackknife variance estimate occurs in the construction of the pseudo-replicates. The method used by NAEP is detailed below. This method is applicable to the estimation of a wide range of statistics, is straightforward in its implementation, and, because adjustments were carried through separately for each replicate, approximately accounts for sources of variability due to nonresponse adjustment and post-stratification. Implementation of this method requires 32 re-computations of the statistic of interest.

Specifically, let $t = t(\underline{y}, \underline{w})$ represent the value of the statistic of interest when it is computed on the full sample using the full-sample weights. (The full-sample weight is the reciprocal of the probability of selection of the student, adjusted for nonresponse and post-stratification.)

The computation of pseudo-replicates of any such statistic involves the use of 32 sets of weights, which we shall refer to as JKWT01 through JKWT32. The set of weights JKWTi is identified with the i^{th} PSU pair and is used to compute t_i, the i^{th} pseudo-replicate of the statistic t. The value of this pseudo-replicate is

$$t \quad = t(\underline{y}, \underline{\text{JKWTi}})$$

which is simply the statistic t re-computed by using the weights JKWTi instead of the full-sample weights (W).

The set of weights for the i^{th} PSU pair, JKWTi, are computed as follows:

(1) Let W_j^B be the base weight for student j. The base weight is the reciprocal of the student's overall probability of selection and is not adjusted by post-stratification or adjusted for nonresponse.

(2) Let

$$W_j^{Bi} = \begin{cases} 0 & \text{if student j is in the first PSU of PSU pair i,} \\ 2W_j^B & \text{if student j is in the second PSU of PSU pair i,} \\ W_j^B & \text{if student j is not in either PSU of PSU pair i.} \end{cases}$$

This set of pseudo-replicate base weights effects the elimination of the first PSU of the pair and replaces it in the sample with the second PSU of the pair.

(3) Adjust the set pseudo-replicate base weights produced by Step 2 for nonresponse and post-stratification by treating them as if they were base weights for the sample. These adjustments take into account the grade/age of the student and the mode of administration. The result is JKWTi.

Because JKWTi is the set of pseudo-replicate base weights after adjustment for nonresponse and post-stratification, the effects of those adjustments on the value of the statistic t are approximately accounted for in the estimate of the variance of t attributed to the i^{th} PSU pair.

As an example, details for the computation of the jackknife variance of a weighted mean follow:

> Let Z_k be the value of some measurement of interest for student k and let W_k be that student's full-sample weight. The statistic of interest is the weighted mean value of Z:
>
> $$t = \sum_{k=1}^{n} W_k Z_k / \sum_{k=1}^{n} W_k \quad .$$
>
> Note that if Z_k can only take values 0 or 1, then t is the weighted proportion receiving a value of 1.
>
> Let W_k^i = value of JKWTi for student k. The pseudo-replicate for the i^{th} PSU pair is
>
> $$t_i = \sum_{k=1}^{n} W_k^i Z_k / \sum_{k=1}^{n} W_k^i \quad ,$$
>
> the jackknife variance of the weighted mean is
>
> $$\hat{Var}(t) = \sum_{i=1}^{32} (t_i - t)^2 \quad ,$$
>
> and the jackknife standard error of the mean is the square root.

13.2.4 The Degrees of Freedom of the Variance Estimate

It is important to have an indication of the number of degrees of freedom to attribute to the jackknife variance estimator $\hat{Var}(t)$. The degrees of freedom of a variance estimator provide information on the stability of that estimator: the higher the number of degrees of freedom, the lower the variability of the estimator. In practical terms, the number of degrees of freedom of the variance estimator corresponds to the number of residual degrees of freedom that can be assumed for inferential procedures.

Note that the jackknife procedure estimates the sampling variability of the statistic by assessing the effect of change in the sample at the paired PSU level. For this reason, the number of degrees of freedom of the variance estimator $\hat{Var}(t)$ will be at most equal to the number of PSU pairs. The number of degrees of freedom equals the number of independent pieces of information used to generate the variance. In the current case, the pieces of information are the 32 squared differences $(t_i - t)^2$, each supplying at most one degree of freedom (regardless of how many individuals were sampled within any PSU).

Increasing the number of individuals sampled within any PSU results in a lower estimate of sampling variability because the within-PSU component is reduced. This, however, does not improve the estimation of the between-PSU component of variability, which depends on the number of PSUs selected. (It does slightly reduce the overall error, however.)

The number of degrees of freedom of the sample variance estimator can be strictly less than the number of PSU pairs. For example, suppose that the statistic t is a mean for some subgroup and no members of that subgroup can come from either PSU in the i^{th} PSU pair. (Examples of such a subgroup are any PSU-level partitioning of the population, such as region.) If the pseudo-replicate weights, JKWTi, had not been adjusted for post-stratification, then since no members of the subgroup come from either member of the PSU pair i, the resulting pseudo-replicate t_i would be identical to the overall estimate t so that $(t_i - t)^2 = 0$. In this case, such a PSU pair imparts no information about the variability of the statistic t and thus contributes zero degrees of freedom to the variance.

However, it is generally the case that t_i does not equal t, even when neither member of the PSU pair i contains observations from the subgroup in question. This is because the pseudo-replicate weights have been adjusted for post-stratification without regard to the grouping and so all weights have been altered. In the instance that neither member of the PSU pair i directly contributes to the estimate of t, the component $(t_i - t)^2$ is measuring the effect of post-stratification on the estimate. While being nonzero, such a component is likely to be smaller in magnitude than the squared difference $(t_k - t)^2$ for any PSU pair k which does contribute to the estimate of t.

In general, the squared difference $(t_i - t)^2$ will be estimating the variance component σ_i^2, say, which is the contribution to the sampling variance of the statistic t which can be attributed to the samples within the i^{th} PSU pair. That is, $\hat{Var}(t)$ is estimating

$$\sum_{i=1}^{M} \sigma_i^2 .$$

If a few of the σ_i^2 are markedly larger than the remainder, as in the above case, then $\hat{Var}(t)$ is predominantly estimating the sum of these larger components which dominate the remaining terms. The effective degrees of freedom of $\hat{Var}(t)$ in this case should be nearer to the number of dominant terms. For a nonlinear estimator, the relationship of the number of degrees of freedom to the contribution that each pair makes to the total estimate of variance is more complicated.

An estimate of the effective number of degrees of freedom for Vâr(t) comes from an approximation due to Satterthwaite (1946) which assumes that the differences $t_i - t$ are independent and approximately normally distributed, with zero means but possible different variances, σ_i^2. Hence the squared differences are each distributed like a chi-square random variable with 1 degree of freedom times a constant, σ_i^2. The Satterthwaite approximation to the distribution of Vâr(t) comes from equating the expectation and variance of Vâr(t) with those of a chi-squared distribution (a constant). Specifically, Vâr(t) is approximately distributed like the constant

$$\sum_{i=1}^{M} \sigma_i^2 \quad \text{times a chi-squared random variable with } df_{eff}$$

degrees of freedom, where df_{eff} is the effective number of degrees of freedom of Vâr(t) defined by

$$df_{eff} = \frac{\left(\sum_{i=1}^{32} (t_i - t)^2\right)^2}{\sum_{i=1}^{32} (t_i - t)^4}$$

which is never larger than 32. (See Cochran, 1977, p. 96 for further discussion.)

13.2.5 Alternative Jackknife Estimators

It should be noted that there are a variety of alternative jackknife estimates of variance available in addition to one given here (see Wolter, 1985).

In particular, two commonly used jackknife estimators are

$$1/2 \left(\sum_{i=1}^{32} (t_i - t)^2 + \sum_{i=1}^{32} (t^*_i - t)^2 \right)$$

and

$$1/4 \left(\sum_{i=1}^{32} (t_i - t^*_i)^2 \right)$$

where t^*_i is an analogous pseudo-replicate to t_i formed by eliminating the second PSU of the pair and double counting the first.

In the case of a linear estimator, all of these methods will produce the same result. Furthermore, in the case of the estimation of sampling variability of a ratio estimate (such as a weighted mean), Monte Carlo experimentation based on the Year 15 National Assessment of Educational Progress Design indicated trivial differences in the three estimates (see Lago, Burke, Tepping, & Hansen, 1985). The ETS estimator $V\hat{a}r(t)$ requires half the computations of the other estimators, at apparently minimal loss (in terms of variability of the variance estimator).

Chapter 13.3

ESTIMATION OF VARIABILITY DUE TO IMPUTATION

Robert J. Mislevy

Educational Testing Service

Potential users of NAEP data should be aware of the special properties of the NAEP database that affect the validity of conventional techniques of statistical inference. Because of the specialized methods used to estimate reading and writing proficiencies in NAEP, the resulting proficiency values have different properties from ordinary test scores. Therefore, standard procedures for statistical inference should not be applied to the NAEP data without modification.

13.3.1 Properties of NAEP Data That Result from Proficiency Estimation Procedures

In conventional applications of item response theory (IRT) scaling, the number of items administered to each respondent is sufficient to obtain a reasonably precise estimate of each individual's proficiency. In NAEP, however, the goal is to estimate group means, rather than individual proficiency values. Some respondents may answer only a few question. Procedures described in detail below are used to estimate a distribution of plausible values for each respondent and to draw values at random from this distribution. The resulting values are appropriate for calculating statistics based on certain groups, but do not represent precise estimates of proficiency for individual respondents.

Use of this method of estimating proficiency results in an increase in the variability of statistics such as means and regression coefficients. Thus, there are two reasons that the standard errors of these statistics are larger than the values that would be obtained with conventional formulas: the use of cluster sampling, which results in non-independent observations; and the use of proficiency estimation methodology that provides consistent estimates of selected group characteristics, but does not yield precise estimates for individual respondents.

Another property of the proficiency estimates based on plausible values is that for some subgroups of respondents, mean proficiencies may be biased. This is explained in Chapters 10.3 and 11.4.

13.3.2 Using Proficiency Values to Estimate Variability

Jackknifing provides a reasonable estimate of uncertainty due to sampling from a finite population when the variable of interest is observed without error from every respondent. As noted in Chapter 10.3, however, some of the key reporting variables in NAEP are not observed without error. Although both reading proficiency and writing proficiency are construed to characterize individual respondents, these proficiencies are not observed directly from any respondent. They are instead inferred imperfectly from responses to a few reading or writing exercises.

Each respondent provides answers to too few cognitive exercises to provide an accurate point estimate of his or her ability. However, as described in Chapter 10.3, it is possible to summarize what is known about the proficiency value θ of respondent i given his or her responses to cognitive exercises (x_i) and background variables (y_i) in terms of a probability distribution $p(\theta|x_i,y_i)$. For computational convenience, these distributions have been approximated by a set of five "plausible values" $\tilde{\theta}_1$ through $\tilde{\theta}_5$, drawn at random from $p(\theta|x_i,y_i)$. They are labeled RDVAL1 through RDVAL5 and WRTVAL1 through WRTVAL5 on the user tape. The spread of these plausible values reflects the uncertainty about the θ value associated with that respondent given the observable variables x and y. The background variables y used in constructing these plausible values are:

(1) age;
(2) grade;
(3) region of the country;
(4) parental education;
(5) sex;
(6) ethnicity; and
(7) size and type of community.

Let $t(\theta,y)$ be a statistic, or a function of the values of θ and y in the sample, estimating a population value T. Examples of statistics t would be weighted means, percentile points, and regression coefficients. If θ were observed directly for sampled pupils, it would be possible to approximate the precision of t through standard methods for survey samples, such as the jackknife technique described above; the result would be, say, $\hat{Var}(t)$. This value addresses uncertainty due to sampling only. Using plausible values, the additional uncertainty incurred when θ is not observed directly can be managed in the following manner:

(1) Using the first vector of plausible values for respondent, RDVAL1, evaluate t as if the plausible values were the true values of θ. Denote the result $\hat{t}_1$.

(2) Using the multiple weight jackknife approach, compute the estimated sampling variance of $\hat{t}_1$, or $\hat{Var}(\hat{t}_1)$ with respect to respondents' first vectors of plausible values.

(3) Carry out steps (1) and (2) for the second through fifth vectors of plausible values, thus obtaining $\hat{t}_u$ and $\widehat{\mathrm{Var}}(\hat{t}_u)$ for $u = 2,\ldots,5$.

(4) The best estimate of t obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$\hat{t}_. = \sum_u \hat{t}_u/5 \quad .$$

(5) An estimate of the variance of $\hat{t}.$ is the sum of two components:

$$\widehat{\mathrm{Var}}(\hat{t}.) = \sum_u \widehat{\mathrm{Var}}(\hat{t}_u)/5 + \sum_u (\hat{t}_u - \hat{t}_.)^2/5 \quad .$$

The first component in $\widehat{\mathrm{Var}}(\hat{t}.)$ reflects uncertainty due to sampling respondents from the population; the second component reflects uncertainty due to the fact that sampled respondents' θ's are not known precisely, but only indirectly through x_i and y_i.

The first component in $\widehat{\mathrm{Var}}(\hat{t}.)$ is attainable by jackknife methods for means as described in the preceding section. Jackknifing could also be applied to more complicated statistics such as regression coefficients.

Computations in this manner of statistics t, involving only writing or reading proficiency, in conjunction with the specific background variables y listed above, provides nearly unbiased estimates of the population values T. Statistics involving proficiency and background variables not listed above are subject to biases, the magnitude of which depend in part on the relationship of the excluded background variables to the included background variables. (See Chapters 10.3 and 11.4 for details.)

13.3.3 Multiple Runs with Different Imputes

Estimating variability requires computing a statistic 165 times, including 33 runs to obtain an estimate and a variance estimate from each of the five sets of plausible values. Because the cost of the full procedure may be prohibitive in many studies, approximate procedures may be used to produce reasonable estimates at lower costs.

One method of reducing costs is to use fewer runs on plausible value sets. A statistic computed from a single set of plausible values has the same expectation as the average of the five, but does not take into account the uncertainty surrounding θ values. Use of at least two, but less than five, sets of plausible values to evaluate a statistic will properly account for this uncertainty and will reduce costs at the same time. The occurrences of "5" in the procedure outlined above would be replaced by a "2", "3", or "4" as appropriate. The resulting decrease in computation is accompanied by a decrease in precision for estimating $\widehat{\mathrm{Var}}(\hat{t}.)$.

A second cost-reducing method produces estimates of sampling variability that are more accurate than those obtained by design effects (see Chapter 14.2) but more variable than those obtained by jackknifing all five pseudo-datasets. This method is to estimate a statistic on each pseudo-dataset (in order to estimate variability due to the latency of proficiency) but compute its jackknife variance on only one pseudo-dataset to estimate sampling variability. This procedure was used by NAEP to produce the "almanacs" of estimated effects (see Chapter 13.4 for a discussion of almanacs).

NOTE: It is not appropriate to average the five plausible values associated with each respondent and analyze those averages. The result of such a computation is not generally equal to the correct value.

Chapter 13.4

USE OF THE NAEP ALMANACS

Rebecca Zwick

Educational Testing Service

The sets of tables summarizing NAEP results are referred to as almanacs. This chapter includes (a) an annotated table listing the NAEP almanacs for Year 15 (1983-84) and for reading and writing trends, (b) a description of the format in which information is presented in these almanacs, and (c) some cautionary notes concerning the interpretation and analysis of weighted means and percentages in NAEP.

13.4.1 Table of NAEP Almanacs

Table 13.4(1) lists the almanacs for NAEP Year 15 and for the reading and writing trends. The information that can be found in these almanacs includes item percents-correct and mean proficiency values for reading and writing, as well as responses to background and attitude items. Almanacs that were created for special studies, such as the BIB/pace bridge study are not included in this table. Note that, except where specified, these almanacs include data for the BIB sample only (see Chapter 5). The Year 15 almanacs correspond to the three grade samples (4, 8, and 11). "Age only" students, that is, those who were age 9, but not in grade 4, age 13, but not in grade 8, or age 17, but not in grade 11 were excluded. The trend almanacs correspond to the three age samples (9, 13 and 17). "Grade-only" students (i.e., those in grades 4, 8, or 11, but not in the designated age groups) were excluded. (See Chapter 4 for a further description of age and grade samples.)

13.4.2 Format of Information Contained in NAEP Almanacs

In the sections below, the format of information in the each of the types of NAEP almanacs listed in the table is described.

13.4.2.1 Type 1: Background and Attitude Items

Each almanac page corresponds to a particular background or attitude item, such as, "How far in school did your father go?" The possible responses to the item are listed across the top of the page. Along the left-hand side of the page is a list of socio-demographic groups, such as

Table 13.4(1)
NAEP Year 15 Almanacs - Dates of Issue and Comments

	Type of Almanac	Grade 4	Grade 8	Grade 11
1.	Background and Attitude Items	10/25/84*	6/11/85	11/16/84*
2.	Reading and Writing Items	3/27/86	3/27/86	3/27/86
3.	Background and Attitude Items with Reading Proficiencies	8/19/85	8/1/85	8/23/85
4.	Reading and Writing Items with Reading Proficiencies	8/19/85	8/1/85	8/23/85
5.	Background and Attitude Items with Writing Proficiencies	12/23/85	12/23/85	1/6/86

NAEP Trend Almanacs

	Type of Almanac	Age 9	Age 13	Age 17
6.	Reading Trend: Mean Proficiencies and Percent at or above anchor points (Years 2, 6, 11, 15; BIB and pace samples merged)	8/7/85	8/7/85	8/7/85
7.	Reading Trend: Percent correct for items common to Years 2, 6, 11, 15 (BIB and pace separately)	7/10/85 13 items	7/10/85 22 items	7/10/85 19 items
8.	Writing Trend: Primary trait and holistic scores for 1 item common to Years 5, 10, and 15 and for 2 items common to Years 10 and 15 (pace sample).	2/12/86	2/12/86	4/3/86

* Standard errors are incorrectly estimated; see Section 13.4.3.

male, female, white, black, or Hispanic. Each line of results in the almanac contains five kinds of information:

- (a) the actual sample size, N, for the group to which that line applies,
- (b) the weighted N, which is the sum of the sampling weights for the group,
- (c) a measure of variability of the weighted N[1] which is enclosed in parentheses following the weighted N,
- (d) the weighted percentages of group members who gave the responses listed horizontally across the page, and
- (e) the standard errors of the weighted percentages, which are enclosed in parentheses following the percentages to which they apply.

13.4.2.2 Type 2: Reading and Writing Items

These almanacs have the same format as the Type 1 almanacs, above. The only difference is that the headings listed across the top of the page represent responses to reading and writing items. In the case of reading items, the correct response choice is indicated by an asterisk.

13.4.2.3 Type 3: Background and Attitude Items with Reading Proficiencies

These almanacs include the same type of information contained in the almanacs of Type 1, as well as an additional line of information for each reporting group. This extra line shows the weighted mean reading proficiency (based on the first plausible value only; see Chapter 10.3) for students who gave each of the possible responses to the background or attitude item on that page. The standard error of the mean is given in parentheses following the mean. (As outlined in Section 13.4.3, the standard error includes a component due to sampling variability and a component due to imprecision of measurement.)

Note that the sample of observations on which these almanacs is based is not identical to that used for the Type 1 almanacs. This is because a subset of students have complete background and attitude data, but do not have proficiency estimates (see Chapter 10.2, on LOGIST).

---

[1] For almanacs issued prior to November 1984, this measure was the standard error of the weighted N. For almanacs issued between November 1984 and January 10, 1985, the measure given is the coefficient of variation (C.V.), which is equal to the standard error of the weighted N, divided by the weighted N. For almanacs issued January 10, 1985, or later, the measure given is the rescaled coefficient of variation, C.V.* = 100 x C.V. The importance of C.V.* is described in Section 13.4.3.

13.4.2.4 Type 4: Reading and Writing Items with Reading Proficiencies

These almanacs are of the same format as those of Type 3, except that the headings listed across the top of the age represent responses to reading and writing items.

13.4.2.5 Type 5: Background and Attitude Items with Writing Proficiencies

These almanacs are of the same format as those of Type 3, except that weighted mean writing proficiencies (based on all five plausible values) rather than reading proficiencies are given.

13.4.2.6 Type 6: Reading Trend: Mean Proficiencies

The first page of the trend tables show weighted reading proficiency means (based on all five plausible values) and standard errors for each of the reporting groups listed along the left side of the page, for each of the assessment years (2: 1970-71, 6: 1974-75, 11: 1979-80, and 15: 1983-84) listed across the top of the page. Subsequent pages give the unweighted number and weighted percentage of students in each cell of the first table, and the weighted percentages of students with reading proficiencies at or above each of the behavioral anchoring points (see Chapter 10.5, on scale definition and behavioral anchoring). The Year 15 data in this almanac are based on the combined BIB and pace samples.

13.4.2.7 Type 7: Reading Trend: Percents Correct

The first page in these almanacs has essentially the same format as the Type 6 almanacs. Instead of a mean proficiency value, however, these almanacs give the weighted mean percent-correct, averaged across the reading items common to the four assessments. The subsequent pages provide this information separately for each item. In these almanacs, Year 15 data are provided separately for the BIB and pace samples.

13.4.2.8 Type 8: Writing Trend

Each page of these almanacs corresponds to a single writing item. Reporting variables and assessment years are listed along the left side of the page. The possible score categories are listed along the top. Each entry in the main body of the table gives the weighted percentage of students in a particular reporting group and assessment year who received the specified standard errors are given. Data are provided for one item common to Years 5, 10, and 15, and two items common to Years 10 and 15. Year 15 data in the Type 8 almanacs are for the pace sample only.

13.4.3 Interpretation and Analysis of Weighted Means and Percentages in NAEP

Weighted proficiency means and weighted percentages of students giving a particular response to a background or attitude item are likely to be used in both descriptive and inferential analyses of the NAEP data. In both cases, the standard errors of these statistics should, of course, be considered. As described in Chapter 13, the standard errors of mean proficiencies are larger than those that would be obtained from conventional formulas for two reasons: First, the use of cluster sampling in NAEP results in non-independent observations. Second, the proficiency estimation methodology used in NAEP provides consistent estimates of selected group-level characteristics, but does not yield precise estimates for individual respondents. In the case of weighted percentages of students, only the first reason applies.

In some cases, the standard errors themselves are poorly estimated, as reflected by large values of C.V.* = 100 x C.V. Westat sampling statisticians suggest the following rule of thumb: If the value of C.V.* for a particular line of an almanac is less than 10, it can be assumed that the standard errors for that almanac line are well estimated; if C.V.* is between 10 and 20, the adequacy of the standard error estimates is in question; if C.V.* is greater than 20, the standard error estimates are unacceptable. Hypothesis tests and confidence intervals should not be computed if C.V.* exceeds 20. (In some NAEP almanacs, values of C.V.* greater than 20 are flagged. Also, as noted in the table, the standard errors in the Type 1 (Background and Attitude) almanacs for Grades 4 and 11 were incorrectly estimated because a replicate weight was wrong (see Chapter 13.2).

Another issue that must be considered in conducting statistical analyses, such as comparisons of means or percentages, is the degrees of freedom. In a complex sample, the degrees of freedom are a function of the number of primary sampling units (PSUs) and strata, rather than the number of observations. In NAEP, the upper bound to the number of degrees of freedom available for an analysis is 32, the number of PSU pairs (which is equal to the number of strata minus the number of PSUs). For a given comparison, the number of available degrees of freedom could be less because only a subset of all PSUs is involved. Further reductions in degrees of freedom may result from inequalities of within-PSU variability. Therefore, in order to avoid Type I errors, a stringent critical value should be used in conducting hypothesis tests. This is especially true when multiple tests are to be performed.[2]

---

[2] As a result of the cluster sampling used in NAEP, the means or percentages for two groups of respondents are not, in general, independent even if the groups do not contain any of the same subjects; instead, they may be positively correlated. The effect of this dependency, which is

Finally, as detailed in Chapters 10.3 and 11.4, mean reading and writing proficiencies may be biased in certain subgroups. If the grouping variable is one of those used in the conditioning procedures described in the above chapters, the mean proficiencies will be virtually unbiased. If the grouping variable was not used in the conditioning, the degree of bias will be a function of the relation between the grouping variable and the conditioning variables.

---

typically ignored, is to reduce slightly the likelihood of a statistically significant result. However, the conservativeness introduced by this dependency is likely to be far outweighed by the increased risk of Type I error that can result from the performance of multiple tests and the overestimation of the degrees of freedom.

Chapter 14

SUPPLEMENTARY STUDIES

Albert E. Beaton

Educational Testing Service

In addition to the processes used to develop the parameter estimates included in the NAEP reports, several other studies have been completed which lend to the credibility and usefulness of the NAEP data. Two of these studies are reported here:

- The validity of the NAEP Year 15 reading and writing assessments. The ETS Standards for Quality and Fairness (1983) require, among other things, that each testing program provide evidence for the validity of the test scores as related to their purpose. The NAEP assessment is, of course, quite different from other ETS programs because the data are not used for--indeed cannot be used for--individual decision-making. However, we have addressed the issue of validity of the measuring instruments in this assessment and the results are reported in Chapter 14.1.

- The design effects of the NAEP data. Because the NAEP sampling design was complex and used various natural clusterings of students, the usual formulas used by standard statistical systems for estimating error variances are strictly inappropriate. A design effect is an estimate of the proportionate increase in error introduced by using standard formulas instead of the jackknife or other appropriate formula. We have examined the design effects to help advise potential secondary users of NAEP data about the likely error in using simpler methods and to offer a computationally simple way of approximating the proper error estimates. The results of this study are reported in Chapter 14.2.

Chapter 14.1

VALIDITY ISSUES IN NAEP: YEAR 15 READING AND WRITING ASSESSMENTS[1]

Rebecca Zwick

Educational Testing Service

In evaluating the adequacy of a cognitive or psychological instrument, the most fundamental question is whether it can provide the basis for valid inferences about the respondent characteristics it claims to measure. The topic of test validity is featured prominently in the recently revised Standards for Educational and Psychological Testing (1985), produced by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (referred to as the Joint Standards below) and in the ETS Standards for Quality and Fairness (1983), which serves as a basis for periodic reviews of ETS programs.

Like most published work on validity, the discussions in these standards manuals focus on testing situations in which the goal is to draw conclusions about individual respondents. Validity must be conceptualized differently in the case of large-scale assessments such as NAEP. The goal of NAEP is to make inferences about groups, rather than individuals. In fact, inferences about individuals are rendered impossible because ETS does not retain the identity of respondents to NAEP items. In addition, the proficiency estimation procedures used in NAEP result in values that are appropriate for calculating statistics based on certain groups of students, but do not represent optimal estimates of reading and writing proficiency for individuals.

Furthermore, the kinds of evidence that can be provided in support of validity differ from the information that is typically available for test validation. It is not generally possible, within the context of NAEP, to collect additional data from NAEP respondents or to conduct supplementary research studies for purposes of investigating validity questions. Also, because precise estimates of individual proficiency are not available in NAEP, estimates of correlations involving NAEP reading and writing skills cannot be obtained in a straightforward manner (see below). This further complicates the validation process, which typically involves examination of the correlations between the measure of interest and other variables.

---

[1] The correlational analyses for this section were performed by Tom Jirele.

These limitations do not, of course, exempt NAEP from the obligation of considering validity issues. It is important to recognize, however, that it is not always desirable, or, indeed, possible, to apply conventional test validation procedures in NAEP.

The ETS Standards manual lists three components of validity: content, construct, and criterion-related validity. In the Joint Standards manual, these three components are described instead as three types of validity evidence. This reflects the view of Messick, who pointed out that "different kinds of inferences from test scores require different kinds of evidence, not different kinds of validity" (1980, p. 1014).

According to the Joint Standards, content-related evidence demonstrates that the sample of items on a test are representative of a specified content domain. Because NAEP is not a test, but a survey that uses multiple matrix sampling, the items received by a given student can not be expected to be a representative sample of the subject matter domain (e.g., reading or writing). However, it is possible to evaluate the representativeness of the total item pool corresponding to each domain by applying essentially the same techniques used in conventional content validation.

Construct-related evidence supports the use of the test score as a measure of the characteristic of interest. This concept must be revised slightly in the case of achievement surveys: In NAEP, the goal of construct validity studies is to determine whether the mean reading and writing scale values for selected sociodemographic groups can be interpreted as measures of reading and writing proficiency. (Alternatively, the validity of inferences based on mean item scores, i.e., the proportions of selected groups who answered individual items correctly, could be considered. Item-level analyses can be useful for research that focuses on specific skills; see Writing: Trends Across the Decade, 1974-84 [Applebee, Langer, & Mullis, 1986a]. The present section is primarily concerned with inferences about overall reading and writing proficiency and focuses, therefore, on the validity of conclusions based on item composites.)

Criterion-related evidence shows that test scores are related to pertinent outcome criteria. For instance, in validating a measure used in employee selection, it is important to demonstrate that test scores are related to job performance. Because NAEP is not a test and is not used to make inferences about individuals, no comparable evidence regarding the NAEP assessment instruments can be provided.

In the following sections, content and construct validity issues are discussed with reference to the NAEP reading and writing assessment.

14.1.1 Content Validity

The NAEP reading and writing exercises for 1983-1984 were the product of an elaborate process of item development, described in detail in Chapter

3 and in two NAEP booklets, Reading Objectives: 1983-1984 Assessment (1984) and Writing Objectives: 1983-1984 Assessment (1982). The process consisted of four phases: (1) development and review of educational objectives in reading and writing, (2) development and review of a pool of items corresponding to these objectives, (3) field testing of prospective items, and (4) final review and item selection. External consultants were extensively involved in developing and reviewing objectives and in writing and reviewing prospective items. In the case of the writing assessment, consultants also participated in the development and review of procedures for scoring the item responses. These consultants included subject area experts, curriculum specialists, classroom teachers, school administrators, and parents. Contributors were chosen to represent a diversity of ethnic groups, community types, and geographic regions. The participation of these consultants was achieved through a series of conferences and mail reviews, coordinated by the NAEP advisory committees for reading and writing.

The items selected in the final phase of the process were required to reflect the educational objectives agreed upon in the first phase and to be consistent with established principles of test development. Among other criteria, they had to be judged free from apparent bias against any sociodemographic group.

In short, a great deal of effort was expended to define reading and writing domains and to ensure that the final pool of NAEP items represented these domains adequately. Although all these NAEP exercises are available for item-level analyses, some items were excluded from the reading and writing proficiency scales because of practical considerations or, in the case of reading items, because they were expected to produce violations of unidimensionality assumptions. Details on the criteria used for including items in the proficiency scales are given in Chapters 10.2 and 11.4. The reading scale included 228 of the 340 items originally designated as reading items; the writing scale included 10 of 22 writing items. The development of proficiency scales facilitated the summarization of reading and writing results and the comparison of results across grades and across assessment years. A drawback of the scaling is that the included items cannot be considered fully representative of the domains that were initially defined.

14.1.2 Construct Validity

14.1.2.1 NAEP Reading and Writing Proficiency Scales

In previous assessments, NAEP results consisted of only the responses to individual items. In contrast, analysis of the 1983-1984 data included the development of reading and writing proficiency scales. In this chapter, the properties of the NAEP reading and writing scales that are relevant to validity assessment are considered. Development of the scales is described in detail in Chapters 10 and 11.

The NAEP reading data, which consisted of dichotomous item responses, were scaled using item response theory (IRT) methods. Specifically, the three-parameter logistic model (Birnbaum, 1968) was applied. Because many students had received only a small number of items, precise estimates of each respondent's proficiency could not be obtained. Instead, estimation procedures that would produce consistent estimates of selected group-level characteristics were used. For each student, a proficiency distribution was estimated, conditional on that individual's item responses and on selected demographic characteristics. Theoretically, these estimated distributions could be used to compute statistics of interest, such as subgroup means, via integration. Because evaluation of the required integrals presents computational problems, the statistics of interest can instead be estimated by making use of "plausible values" selected at random from each respondent's distribution. For each respondent, five plausible values were drawn, each of which can be viewed as an estimate of that student's unknown proficiency value. This proficiency estimation methodology serves the goals of NAEP in that, unlike conventional IRT methods, it provides consistent estimates of group characteristics when applied to the sparse data available for individual NAEP respondents. A drawback is that the plausible values are not optimal estimates of individual proficiency.

Development of the writing proficiency scale presented an additional challenge: unlike the reading items, which were dichotomously scored, the writing items were scored on a five-point scale. (Only the primary trait scores were used for the proficiency scale; see Chapter 13.4.) Although generalizations of IRT methods have been developed for rating scale data, application of these scaling techniques to the NAEP writing data did not produce satisfactory results. Therefore, a multiple regression approach, called the average response method (ARM), was used to develop the writing proficiency scale.

Although students had answered different subsets of the ten NAEP writing items selected for inclusion in the scale, it was possible, because of BIB spiralling (see Chapter 4), to obtain the matrix of pairwise correlations between the items. Therefore, for each respondent, a predicted mean score on the ten items could be derived, conditional on that individual's item responses and on selected demographic characteristics. A writing plausible value, analogous to the reading plausible value was then obtained by adding to this predicted score a random term representing the uncertainty of the respondent's predicted mean, given that individual's demographic characteristics and item responses.

Although the methods of scale development were not identical for reading and writing, both methods yield so-called plausible values, which are not optimal estimates of individual proficiency. The scale construction methodology has important implications for the assessment of validity in NAEP. For example, as a result of the scaling approach, sample means for certain NAEP subpopulations are biased. Also, correlation coefficients based on plausible values yield seriously attenuated estimates of the relations between NAEP reading proficiency, writing proficiency, and other variables of interest. These issues are addressed in the following sections.

14.1.2.2 Validity Evidence Based on Group Differences in Mean Proficiency

NAEP has focused its energies on the calculation of means and standard errors for subpopulations that are of primary interest. These selected groups are based on the following demographic variables: grade (4, 8, or 11), sex (male or female), ethnicity (white, black, Hispanic, or other), size and type of community (advantaged urban, disadvantaged urban, or other), region (northeast, southeast, central, or west), and parental education (did not graduate from high school, graduated from high school, received schooling beyond high school, or unknown). (See Chapter 12 for detailed definitions of these variables.) These demographic characteristics were used as conditioning variables for developing both the reading and writing scales. (An additional variable, grade/age status, was used in developing the reading scale. See Chapter 10.3.) Because of their inclusion in the conditioning, estimates of proficiency means for subpopulations based on these variables are virtually unbiased. The means of subgroups other than the primary demographic groups are biased to varying degrees, as explained in Section 10.3.5 and cannot, therefore, provide the basis for strictly valid inferences. (It should be noted that the problem of bias does not in any way affect inferences based on item-level data.)

In investigating the construct validity of NAEP, it is important to determine whether the patterns of proficiency means for the primary demographic groups are consistent with educational theory. As a rather obvious example, the mean for Grade 11 is expected to be highest, followed by Grades 8 and 4 in that order. Of course, confirmation of this expectation is not sufficient to demonstrate validity; disconfirmation, however, would cast serious doubt on the results. In the following section, details are provided on grade differences and other group differences about which theory-based hypotheses could be formed. The mean proficiency values on which this section is based, along with their standard errors, are given in Table 14.1(1). (The sources of these data are The Reading Report Card: Progress Toward Excellence in Our Schools [1985], and The Writing Report Card: Writing Achievement in American Schools, 1984 [Applebee, Langer, & Mullis, 1986b].)

Grade. For both reading and writing, the proficiency means for the three grades are appropriately ordered. In both subject areas, the difference between Grades 8 and 11 is smaller than the difference between Grades 4 and 8. This is consistent with expectations for two reasons: First of all, the Grade 8 and 11 students are only three grades apart, whereas the Grade 4 and 8 students are four grades apart. Also, theories of cognitive development predict greater improvement in reading and writing proficiency between Grades 4 and 8 than in the teenage years.

Table 14.1(1)

Reading and Writing (ARM) Proficiency Means for Selected Groups
(standard errors in parentheses)

	Grade 4		Grade 8		Grade 11	
	Reading	Writing	Reading	Writing	Reading	Writing
Total	217.5(.7)	1.58(.01)	260.7(.5)	2.05(.01)	289.3(.8)	2.19(.01)
Sex						
Male	215.1(.9)	1.50(.01)	257.0(.6)	1.96(.01)	284.5(1.0)	2.09(.01)
Female	220.0(.7)	1.66(.01)	264.5(.6)	2.14(.01)	294.3(.9)	2.29(.01)
Parental Education						
Did not graduate H.S.	200.2(1.2)	1.43(.03)	244.2(.7)	1.89(.02)	269.5(1.2)	1.99(.02)
Graduated H.S.	215.5(.8)	1.54(.01)	255.5(.7)	2.02(.01)	281.8(.7)	2.15(.01)
Post H.S.	227.4(1.1)	1.66(.01)	271.8(.7)	2.13(.01)	300.6(.9)	2.27(.01)
Size/Type of Community						
Rural	207.7(2.5)	1.53(.02)	259.9(2.3)	2.03(.03)	284.6(3.2)	2.13(.03)
Disadvantaged urban	198.5(1.5)	1.42(.02)	241.6(1.9)	1.88(.02)	267.8(2.5)	2.01(.02)
Advantaged urban	234.5(2.3)	1.70(.02)	276.9(2.6)	2.21(.02)	300.2(3.0)	2.28(.02)

Sex. Research in cognitive development has consistently demonstrated that the verbal skills of girls are superior to those of boys. This held true for the NAEP results in both reading and writing. The magnitude of the superiority increased slightly with grade level, which is consistent with the findings of some prominent researchers (see Jacklin, 1979; Mussen, Conger, & Kagan, 1969).

Parental Education. Previous research has consistently shown that student achievement tends to be highest for those whose parents have received the most education (e.g., Jones, Burton, & Davenport, 1982). The NAEP results provided an additional confirmation of this for both reading and writing, within each grade. It should be noted that the NAEP data on parent education are based on students' reports. The category definitions for parental education (given in abbreviated form in Table 14.1(1)) are: (1) neither parent graduated from high school, (2) at least one parent graduated from high school (but neither parent received post-high school education), and (3) at least one parent received some post-high school education.

Size and Type of Community. NAEP reading and writing results were reported for three community types of special interest, defined as follows:

> Rural communities: Students in this group attend schools in areas with a population under 10,000 where many of the residents are farmers or farm workers.
>
> Disadvantaged urban communities: Students in this group attend schools in or around cities having a population greater than 200,000 where a high proportion of the residents are on welfare or are not regularly employed.
>
> Advantaged urban: Students in this group attend schools in or around cities having a population greater than 200,000 where a high proportion of the residents are in professional or managerial positions.

(Note that only about a third of the NAEP respondents lived in communities that fell into one of these categories.) As would be expected, based on these definitions, achievement was higher for advantaged urban students than for disadvantaged urban students in all three grades. Expectations for rural students were less clear. In fact, their achievement levels were consistently above the disadvantaged urban group, but below the advantaged urban students.

14.1.2.3 Validity Evidence Based on Correlations with Attitude Variables and PSAT Scores

Construct validity investigations typically include examination of the correlations between test scores and other variables of interest to determine whether they are consistent with a hypothesized pattern. For instance, scores on a reading test would be expected to correlate more highly with other reading measures than with scores on a math test. In NAEP, however, conventional reading and writing scores are not available. If the plausible values are used in computing correlation coefficients, the correlation estimates will be severely attenuated because the plausible values do not represent precise estimates of proficiency for individual respondents. Furthermore, there is no straightforward way to achieve a satisfactory correction for attentuation. (In principle, plausible values could have been constructed to take into account the joint marginal characteristics of reading and writing. This would have prevented the attenuation problem. However, practical considerations necessitated that plausible values be constructed separately for reading and writing.) Therefore, an alternative approach, described below, was used to estimate correlations between reading, writing, and other variables of interest.

As an illustration of the method, consider the correlation between reading and writing skills, as assessed in NAEP. If all N students had answered each of the r reading items and each of the w writing items, the data could be represented as a $N \times (r + w)$ matrix, X. It would be possible to obtain a total score on reading (i.e., the number of reading items answered correctly) and a total score on writing. The crossproducts matrix corresponding to the total score for reading and the total score for writing could be computed in one of two ways:

(1) First compute $Y = XT$, where T is a $(r + w) \times 2$ transformation matrix that sums the reading and writing item scores. The first column of T has r ones, followed by w zeroes. The second column has r zeroes, followed by w ones. The $N \times 2$ matrix Y contains the reading and writing total scores for each of the respondents. Then compute $C = Y'Y$, the 2×2 crossproducts matrix of the total scores on reading and writing.

(2) Alternatively, start by computing $A = X'X$, the $(r + w)$-dimensional crossproducts matrix of the reading and writing item scores. Then obtain C by computing the matrix product $T'AT$.

In either of these methods, the matrix of correlations for the reading and writing total scores could be obtained through a transformation of C.

Because the complete data matrix X is not available in NAEP, method 1 cannot be applied. However, an approach similar to method 2 can be used. Although $A = X'X$ cannot be computed, BIB spiralling allows the computation of the matrix A^* of pairwise crossproducts of the $r + w$ items. By substituting A* for A, the approach in method 2 can be applied. This is the approach used in the analyses described below. A subroutine for obtaining correlation matrices from incomplete data was first applied, followed by the Transform Cross Products Matrix (TCM) algorithm developed by Beaton (1964). Because most of the missing data in NAEP are a result of the BIB design and can therefore be treated as missing at random, the resulting correlations can be interpreted in essentially the same manner as correlations between total test scores in the complete data case. (Because conventional reliability formulas do not apply in the present context, attempts to correct these TCM correlations for attenuation may lead to misleading results. Therefore, the reported correlations have not been corrected for attenuation. The degree of attenuation of correlations computed using the TCM approach, however, is much less severe than the attenuation that results when correlations are computed using the plausible values.)

All reading and writing items used in the NAEP proficiency scales were included in the correlational analyses, provided that they had been BIB-spiralled with all other items. For reading items, this criterion resulted in the use of 108 out of 118 calibrated items for Grade 4, 106 out of 124 for Grade 8, and 95 out of 113 for Grade 11. In the case of writing items, all items used in the ARM scaling were included: 8 items for Grade 4, 10 for Grade 8, and 6 for Grade 11. Thus, although the resulting correlations do not directly reflect the properties of the reading and writing scales, they are based on nearly the same items.

Only a limited number of variables were available for correlational analysis of the NAEP data. Of primary interest was the correlation between reading and writing. This correlation was expected to be moderately high. In addition, 16 background and attitude items were selected for inclusion in the correlational analysis. These items, which were administered using a multiple choice format, included questions about the language spoken in the student's home, the grades received by the student, the student's perceptions of his or her ability in reading and writing, and the amount of reading and writing done by the student, both in and out of school. Scores on these attitude variables were not summed; instead, the correlation of each item with the reading and writing was considered separately. For purposes of these analyses, responses to the background and attitude items were re-coded in such a way that their correlations with reading and writing were expected to be positive. For example, on items that asked students how often they read stories or novels in their free time, responses of "almost every day" received the highest numerical code; responses of "never or hardly ever" received the lowest. These correlations were expected to be of low to moderate size.

Finally, for a subset of Grade 11 students, verbal and quantitative PSAT scores were available. These were obtained without violating confidentiality requirements. Lists of PSAT scores for all PSAT takers within a school were provided by ETS to schools participating in NAEP. If students selected for the NAEP sample had taken the PSAT, school personnel entered the scores onto the students' NAEP records. ETS did not retain any information about the identity of the NAEP participants.

Ideal evidence for construct validity would be a finding that correlations of reading and writing with PSAT verbal scores were quite high, exceeding the correlation between PSAT verbal and quantitative scores, whereas correlations of reading and writing with the PSAT quantitative scores were only moderate. Results of this kind could be considered as informal evidence of both convergent and discriminant validity (Campbell & Fiske, 1959). Convergent validation shows that the measure of interest is highly correlated with independent measures of similar constructs. Discriminant validation demonstrates that the measure under evaluation is not highly correlated with variables from which it is theoretically expected to differ.

14.1.2.4 Correlations of Reading, Writing, PSAT Scores, and Selected Background Variables

Separate correlational analyses were conducted for each grade and for the subsample of eleventh graders who took the PSAT. These analyses are reported in Tables 14.1(2), (3), and (5). Table 14.1(2) shows, for each grade, the correlation matrix for reading, writing, and four of the sixteen background and attitude variables included in the analysis: language spoken in the home, grades in school, and student self-assessments of ability in reading and writing, respectively. A correlation matrix for these six variables, as well as the PSAT verbal and quantitative scores is given in Table 14.1(3) for the PSAT subsample. The precise definitions of all variables used in these analyses are given in Table 14.1(4). Table 14.1(5) gives the correlations of reading and writing with the remaining twelve background and attitude items, which concern frequency of reading and writing activities, for Grades 4, 8, and 11, and for the PSAT subsample. The item texts and response codes for these twelve items are given in Table 14.1(6).

The correlational analyses were based on approximately 20,000 unweighted observations in each of Grades 4, 8, 11, and 8,500 observations in the PSAT subsample. Because of BIB spiralling, however, the number of respondents available for the estimation of each correlation in the original matrix was much less. For Grades 4, 8, and 11, most correlations were based on 200 to 300 observations; for the PSAT subsample, most correlations were based on about 100 observations.

The correlation between reading and writing, which was of primary interest, was .64 for Grade 4, .60 for Grade 8, .51 for Grade 11, and .53 for the PSAT subsample. The size of these correlations is generally consistent with expectations, although the considerably lower correlations for Grade 11 and for the PSAT subsample were somewhat surprising. (Standard errors of these correlations cannot be obtained by standard methods. Jackknifed standard errors could be computed, but are not currently available.) It is likely that the smaller correlations result from greater homogeneity among Grade 11 NAEP participants than among the fourth or eighth graders. The variability of number-right scores for reading and writing can be estimated using the TCM method described above. For Grade

Table 14.1(2)
Correlations of Reading, Writing, and Selected Background Variables*

Grade 4

	Reading	Writing	Home Language	Grades	Kind of Reader	Kind of Writer
R	1.00					
W	.64	1.00				
HL	.15	.12	1.00			
G	.39	.28	.07	1.00		
KR	.31	.19	.05	.21	1.00	
KW	.02	-.09	-.01	.16	.22	1.00
Grade 8						
R	1.00					
W	.60	1.00				
HL	.11	.10	1.00			
G	.47	.33	.02	1.00		
KR	.33	.23	.08	.26	1.00	
KW	.17	.15	.04	.20	.27	1.00
Grade 11						
R	1.00					
W	.51	1.00				
HL	.16	.09	1.00			
G	.43	.32	.02	1.00		
KR	.33	.18	.08	.24	1.00	
KW	.29	.25	.09	.27	.31	1.00

*Variables are defined in Table 14.1(4). Methodology used to obtain correlations is described in the text.

Table 14.1(3)

Correlations of Reading, Writing, PSAT Scores, and
Selected Background Variables*
(PSAT subsample only)

	Reading	Writing	PSAT-V	PSAT-Q	Home Language	Grades	Kind of Reader	Kind of Writer
R	1.00							
W	.53	1.00						
PV	.67	.32	1.00					
PQ	.57	.26	.67	1.00				
HL	.20	.10	.15	.08	1.00			
G	.50	.35	.46	.50	.05	1.00		
KR	.29	.29	.38	.17	.13	.24	1.00	
KW	.31	.28	.37	.22	.09	.32	.53	1.00

*Variables are defined in Table 14.1(4). Methodology used to obtain correlations is described in the text.

Table 14.1(4)

Definition of Variables for Analyses in Tables 14.1(2) and 14.1(3)

Reading: All calibrated reading items were included, provided that they were BIB-spiralled with all other items (see Chapter 5). This criterion resulted in the use of 108 out of 118 calibrated items for Grade 4, 106 out of 124 for Grade 8, and 95 out of 113 for Grade 11.

Writing: All writing items used in the ARM scaling were included (8 items for Grade 4, 10 items for Grade 8, and 6 items for Grade 11.)

Home Language: What language do you speak most often in your home?

1 = English
0 = Other

Grades: Which of the following best describes your grades so far in school?

9 = Mostly A (a numerical average of 90 - 100)
8 = Both A and B
7 = Mostly B (80-89)
6 = Both B and C
5 = Mostly C (70-79)
4 = Both C and D
3 = Mostly D (60-69)
2 = Both D and E
1 = Mostly below D (below 60)

Kind of Reader: What kind of reader do you think you are?

3 = A very good reader
2 = A good reader
1 = A poor reader

Kind of Writer: [Instructions - The following sentences are true for some people. They may or may not be true for you, or they may be true for you only part of the time. How often is each of the following sentences true for you?] I am a good writer.

5 = Almost always
4 = More than half the time
3 = About half the time
2 = Less than half the time
1 = Never or hardly ever

PSAT-V: Score on verbal section of Preliminary Scholastic Aptitude Test

PSAT-Q: Score on quantitative section of Preliminary Scholastic Aptitude Test

Table 14.1(5)

Correlations of Reading and Writing with Frequency of Reading and Writing Activities*

	Grade 4		Grade 8		Grade 11		PSAT Subsample	
Reading Activities	R	W	R	W	R	W	R	W
1. During free time, how often do you read a book?	.17	.20	.19	.23	.18	.07	.19	.02
2. During free time, how often do you read a newspaper or magazine?	.02	.09	.12	.11	.17	.16	.17	.10
3. How many pages do you read in school and for homework?	.09	.06	.10	.10	.21	.16	.16	.12
4. How often do you read a story or novel?	.20	.06	.37	.26	.27	.17	.31	.10
5. How often do you read a newspaper?	.10	-.06	.17	.22	.17	.15	.09	.03
6. How often do you read a magazine?	.02	-.02	.13	.13	.09	.07	.11	.00
7. How often do you read for fun on your own time?	.23	.11	.30	.28	.22	.10	.22	.21
8. How often do you read on your own in school?	.21	.16	.14	.24	.12	.14	.15	.04
Writing Activities								
9. How much of English class is spent learning to write?	-.22	-.10	-.01	.02	-.07	.07	.08	.11
10. How many stories did you write for English last week?	-.11	-.08	-.13	-.06	-.23	-.11	-.20	-.07
11. How many writings did you do last week that were not for school?	.06	.01	.09	.03	.08	.03	.02	.11
12. How often do you write stories or poems that are not for school?	-.14	-.11	-.12	-.02	.00	.02	.02	-.12

*Item texts and response codes are given in Table 14.1(6). Methodology used to obtain correlations is described in the text.

Table 14.1(6)

Item Text and Response Codes for Reading and Writing

When you have free time, how often do you do each of the following?

1. Read a book
2. Read a newspaper or magazine

3 = Every day or almost every day
2 = About once a week
1 = Once a year or less

3. About how many pages a day do you have to read in school and for homework?

5 = More than 20
4 = 16 - 20
3 = 11 - 15
2 = 6 - 10
1 = 5 or less

How often do you read each of the following?

4. Part of a story or novel
5. A newspaper
6. A magazine

5 = Almost every day
4 = Once or twice a week
3 = Once or twice a month
2 = A few times a year
1 = Never or hardly ever

How often do you do each of the following things?

7. Read for fun on your own time
8. Read on your own in school

5 = Almost every day
4 = Once or twice a week
3 = Once or twice a month
2 = A few times a year
1 = Never or hardly ever

Table 14.1(6)
(continued)

9. About how much of your time in English class is spent learning to write?

5 = Most of the time
4 = More than half the time
3 = About half the time
2 = Less than half the time
1 = None or almost none of the time

About how many of each of the following kinds of writing did you do for your English class last week?
10. A story

3 = 3 or more
2 = 1 or 2
1 = None

11. About how many times during last week did you write something that was NOT a school assignment?

4 = 3 or more
3 = 2
2 = 1
1 = None

How often do you write each of the following things?
12. Stories or poems that are not schoolwork

4 = Almost every day
3 = Once or twice a week
2 = Once or twice a month
1 = Never or hardly ever

11, the standard deviations were 15.0 and 2.6 for reading and writing, respectively; for the PSAT subsample, the corresponding values were 10.9 and 2.6. The standard deviations for Grades 8 and 4 were substantially larger: 16.9 and 3.7 for Grade 8 and 21.2 and 3.6 for Grade 4. The smaller variability in Grade 11 probably occurred in part because some low achievers drop out of school before Grade 11. In addition, the rate of participation in NAEP was somewhat smaller for Grade 11 students than for students in Grades 4 or 8, which could further restrict the range of reading and writing proficiency in the Grade 11 sample. Because only a select subgroup of students take the PSAT, the variability for this subsample was still smaller.

The findings based on PSAT scores were moderately supportive of the validity of the NAEP reading and writing assessments. Reading had quite a high correlation, .67, with PSAT verbal scores, and a lower correlation, .57 with PSAT quantitative scores. Similarly the correlation between writing and PSAT verbal scores was .32, which was higher than the correlation of .26 between writing and PSAT quantitative scores. However, the large correlation of reading and PSAT-V was matched in size by the correlation of PSAT-V with PSAT-Q. Also, the correlation of .57 between reading and PSAT-Q was slightly larger than the correlation between reading and writing, which was .53. In interpreting these patterns of correlations, it is necessary to keep in mind that the small number of writing items and the consequent low reliability result in the attenuation of the correlations of writing with other variables.

The correlations of reading and writing with the background and attitude items included in Tables 14.1(2) and (3) were, for the most part, small to moderate positive correlations. In general, reading had higher correlations with these items than writing. This was not unexpected, given that there were many more reading than writing items, resulting in higher reliability. The correlations of reading and writing with home language were small, ranging from .11 to .20 for reading and .09 to .12 for writing. Correlations with self-reported grades in school were moderate, ranging from .39 to .50 for reading and from .28 to .35 for writing. The question, "What kind of reader do you think you are?" had correlations ranging from .29 to .33 with reading and .18 to .29 with writing. The question asking the students about their writing ability had correlations ranging from .17 to .31 with reading and from .15 to .28 with writing in the Grade 8, Grade 11, and PSAT subsamples. However, in Grade 4, the corresponding correlations were .02 and -.09. A clue to this discrepancy is provided in Table 14.1(4): The instructions and phrasing of the question are probably confusing to fourth graders. In all three grades, the correlation of the "kind of writer" question with reading was slightly higher then the correlation with writing. This may result from the greater number of reading items.

Results for the twelve additional background and attitude items included in the analysis are shown in Table 14.1(5). These items pertained to the frequency with which students participated in reading and writing activities, both in and out of school. For items pertaining to reading activities, most correlations were small and positive. The items that were most highly correlated with reading and writing pertained to the frequency of reading books during free time, reading stories or novels, reading for fun, and reading on one's own in school. Most of the reading activity items were more highly correlated with reading than with writing. This was expected, both because the items referred to reading activities and because more reading than writing items were included in the analysis.

Results for the items pertaining to writing activities were much more puzzling. For items 9, 10, and 12, most of the correlations with reading and writing were negative. In considering some related results, the authors of The Writing Report Card speculated that English teachers may assign more writing to low-achieving students than to skilled students. This would not, however, explain the negative correlations for item 12, which refers to writings that are not schoolwork. It could be hypothesized that there is a tendency for low achievers to overstate their accomplishments, resulting in negative associations of proficiency with self-reported frequency of reading and writing. This still would not account for the fact that item 11, which appears to be similar to item 12, has positive correlations with reading and writing.

The analyses of Tables 14.1(3) and (5) were repeated within subsamples based on sex and ethnicity. For Grade 8, analyses were conducted separately for males and females. Both this analyses and previous correlational analyses for all three grade samples showed that the correlational structure for males and females was essentially the same. For each of the three grades, correlational analyses were also conducted separately for whites, blacks, and Hispanics. At each grade level, the unweighted sample size for these correlational analyses was 13,000 to 17,000 for whites, slightly over 3,000 for blacks, and 2,000 to 3,000 for Hispanics. The typical number of respondents available to estimate each coefficient in the original correlation matrices was about 150 to 200 for the white samples, about 35 for the black samples, and about 25 for the Hispanic samples.

Although some ethnic group differences were evident, there were few consistent or interpretable patterns. One group difference that did show some consistency involved the correlations between reading and writing, displayed below.

	Grade 4	Grade 8	Grade 11
Hispanic	.66	.70	.54
White	.61	.57	.49
Black	.64	.48	.37

In Grades 8 and 11, the correlation for Hispanics was higher than the correlation for whites, which, in turn, was considerably higher than that for blacks. In Grade 4, all ethnic groups had similar correlations. This finding may be attributable in part to differences across ethnic groups in the range of reading and writing proficiency. Variability in reading and writing tended to be somewhat higher in Hispanics than in blacks and whites. Another finding that was consistent across grades was the ethnic group differences in the correlations of reading activity items (1-8 on Table 14.1(5)) with reading and writing. These correlations tended to be smaller for blacks than for whites and Hispanics. Also, the correlations of items 9, 10, and 12 with reading and writing tended to be somewhat more negative for blacks and Hispanics than for whites.

It is difficult to assess the relevance of the correlational analyses of Table 14.1(5) to the validity of the NAEP reading and writing assessment because the validity of the responses to the reading and writing activity items is itself in question. That is, no data are available on the relation between these student reports and the actual frequency of reading and writing activities. Therefore, correlations based on these items (and on the self-reported items included in Table 14.1(3)) can not be given as much weight in the validity assessment as the content-related evidence, the correlations between reading, writing, and PSAT scores, and the patterns of proficiency means across sociodemographic groups.

14.1.3 Summary of Validity Evidence

The NAEP reading and writing exercises were the products of an elaborate item development process, which included the participation of subject area experts, curriculum specialists, teachers, school administrators and parents, in addition to the NAEP committees on reading and writing. The process involved developing educational objectives, preparing items corresponding to these objectives, field testing the items, and finally, selecting items for inclusion in the assessment. The final set of items was judged to be consistent with the specified content domains and to conform to established principles of test development. All these assessment items are available for item-level analyses. However, because of practical and methodological considerations, only a subset of the items was included in the reading and writing proficiency scales. Therefore, these scales do not fully correspond to the established reading and writing domains.

In general, the construct-related evidence based on analysis of proficiency means and correlations appears to support the validity of the NAEP reading and writing assessments. Differences across sociodemographic groups based on grade, sex, parental education, and size and type of community tended to be consistent with findings of previous research and with theories of cognitive development.

Correlational analyses of reading, writing, PSAT scores, and selected background and attitude items produced an unexpected finding: Three of four items pertaining to frequency of writing activities were negatively correlated with reading and writing. However, most of the correlational results were quite supportive of the validity of the NAEP reading and writing assessment. The correlation between reading and writing was moderately high, as expected, and that reading and writing were both more highly correlated with the PSAT verbal scores than with the PSAT quantitative scores. Reading and writing had moderate correlations with self-reported grades and small to moderate correlations with student self-assessments of reading and writing ability.

Chapter 14.2

DESIGN EFFECTS[1]

Eugene G. Johnson

Educational Testing Service

The major computational load in measuring uncertainty for any statistic is in the estimation of the uncertainty due to sampling variability. The jackknife variance estimation procedure requires that the statistic be repeatedly recomputed to obtain an estimate of the sampling variance of the statistic. In the current design, this involves 33 computations, once for the overall estimate and once for each of the 32 PSU pairs. If the population value of interest is based on proficiency values, so that the statistic is computed on a set of plausible values, then, for reasons given in Chapter 10.3, the entire process should be repeated once for each set of plausible values.

This section describes how to approximate the sampling variability of any statistic by less computationally intensive methods. In the case that the statistic is based on proficiency values, the method will provide an estimate of the sampling variability for the statistic. The component of variability due to imputation should still be estimated by recomputing the statistic on different sets of plausible values (see Section 13.3.3).

It is inappropriate to estimate the sampling variability of any statistic based on the NAEP database by using simple random sampling (SRS) formulas. These formulas, which are the ones used by most standard statistical software such as SPSS and SAS, will produce variance estimates which are generally much smaller than is warranted by the sample design.

It may be possible to account approximately for the effects of the sample design by using an inflation factor, the design effect, developed by Kish (1967) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the conventional variance estimate based on a simple random sample with the same number of elements. To avoid sources of bias due to improper representation, this conventional estimate must use the sampling weights. The design effect may be used to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the design. In practice, this is often accomplished by dividing the total sample size by the design effect and using this effective sample size in the computation of errors. Note that the value of the design effect depends on the type of statistic computed and the variables considered in a particular analysis as well as the combined clustering, stratification, and weighting effects occurring among sampled elements.

---

[1] The statistical programming for this section was provided by Bruce Kaplan, David Freund, and Laurel Barnett. The figures were produced by Ira Sample.

Based on empirical results and theoretic considerations, Kish and Frankel (1974) have developed several conjectures about design effects:

(1) Generally, the design effects for complex statistics from complex samples are greater than one, causing variances based on simple random sampling assumptions to tend to be underestimates.

(2) The design effects for complex statistics (such as regression coefficients) tend to be smaller than the corresponding design effects for means of the same variables. Hence, the design effects for means, which are more easily computed, tend to give overestimates of the design effects of complex statistics.

(3) Qualitatively and comparatively, the design effects of complex statistics tend to resemble those of means; variables with a high design effect of the mean also tend to have high design effects for complex statistics involving those variables.

To incorporate the design effect idea in a statistical analysis, proceed in the following manner:

(1) For a given class of statistics (e.g., means, percentile points, regression coefficients), compute the jackknife variance for a number of cases corresponding to the estimate of a particular statistic from a specified subgroup of the population. The cases should cover the range of situations for which the approximation is to be used. If various subpopulations are to be considered, it is important to have information on the relative variability within each subgroup. This is especially important if certain subgroups are more highly clustered in the sample.

(2) For the identical cases, compute the conventional estimate of the variance. This estimate must take the sample weights into account to avoid problems of bias due to improper representation. To account properly for the difference between the number of individuals being sampled and the total of the sampling weights, the weights should be scaled so that their sum equals the sample size.

(3) For each case, compute the design effect where the design effect for case j is

$$\text{deff}_j = \text{Var}_{JK}(t_j)/\text{Var}_{CON}(t)_j$$

the ratio of the jackknife variance estimate of the statistic to its conventional variance estimate.

(4) If the design effects for the various cases are tolerably similar, choose an overall composite design effect. If the design effects for certain subgroups appear to cluster around a markedly different value from the remaining cases, treat those subgroups separately.

(5) In the case that a consistent overall design effect has been found:

  (a) rescale the weight of each individual so that the sum of the scaled weights is equal to the effective sample size

$$N_{eff} = \frac{\text{sample size}}{\text{design effect}}$$

  (b) conduct a traditional weighted analysis using these scaled weights

(6) The degrees of freedom for any variance estimates obtained by using this approach is still at most 32, the number of PSU pairs, as it was for the jackknife. Accordingly, tests of significance produced by standard programs (which will use, for the error degrees of freedom, the effective sample size minus the number of parameters) should be interpreted with extreme caution because they are likely to be too liberal. Significance and inferential procedures are properly based on the smaller error degrees of freedom (32).

14.2.1 Some Design Effects from the Year 15 Reading Assessment

As an example of the distribution of design effects to be expected from NAEP data, we consider the design effect for the key statistic, P, the estimated proportion of a specified subgroup of the population who would correctly respond to a given assessment exercise. This estimate, which is a weighted mean of the responses of individuals in the subgroup to the exercise (where an individual's response is either 0 or 1), has a design effect of the form

$$\text{deff}(P) = \text{Var}_{JK}(P)/(P(1 - P)/N)$$

In the above, N is the total number of individuals in the subgroup responding to the exercise, $\text{Var}_{JK}(P)$ is the jackknife variance of P, and P(1 - P)/N is the conventional variance estimate of P. (Although the estimate P(1 - P)/N has the same form as the simple random sampling estimator of the variance of P, the sample weights have been taken into account via the weighted estimation of P.)

The distributions of design effects for proportions correct by grade and by demographic subgroup within grade across all cognitive reading itemspresented in the Year 15 assessment are indicated in Figures 14.2-1a through 14.2-3c, and Tables 14.2(1) through 14.2(3).

Table 14.2(1) addresses the distributions of the design effects for the 131 cognitive reading exercises presented to Grade 4 students as whole ("total") as well as for a variety of demographic subgroups: sex; race/ethnicity (White, Black, Hispanic, other); region (Northeast, Southeast, Central, West); parental education (At Most High School, Graduated High School, Post-High School, Unknown); and Size and Type of Community (Rural, Low Metropolitan, High Metropolitan, Big City, Urban Fringe, Medium, City, Small Place). For each of these groupings of Grade 4 students, Table 14.2(1) provides the lower quartile (LoQ), median, upper quartile (HiQ) and maximum design effect as well as the mean design effect and the percent of design effects less than 2 and 2.5.

A graphical display of the distributions of the design effects for the same sets of students appears as the boxplots (strictly, box-and-whiskers plots) shown in Figures 14.2-1a through 14.2-1c. The left and right margins of the box in each boxplot correspond to the lower and upper quartiles, the vertical line within the box to the median; the minimum and maximum values are indicated by the ends of the horizontal lines (see Tukey, 1977 for further details). Because the distributions of the design effects are badly skewed, the plots were symmetrized by plotting the log (base 10) of the quantiles of the design effects.

Equivalent information on the distributions of design effects for the 130 cognitive reading exercises presented to Grade 8 students appears as Table 14.2(2) and Figures 14.2-2a through 14.2-2c. The 116 cognitive reading items presented to Grade 11 students are addressed by Table 14.2(3) and Figures 14.2-3a through 14.2-3c.

The particular demographic variables shown (sex, race/ethnicity, region, parental education, and size and type of community) were selected because (1) they are major variables in NAEP reports and (2) they reflect different types of divisions of the population which might have different levels of sampling variability.

The tables and figures show that the design effects are predominantly larger than 1, indicating that standard variance estimation formulas will be generally too small, sometimes markedly so. Further, the distributions of design effects appear different for certain subgroups of the population.

A striking feature of the tables is the apparent lower sampling variability of the Grade 8 data relative to the other two grades. In nearly every case, the median design effect for a subgroup based on Grade 8 data is smaller than the equivalent medians for the other two grades. (This is also true for the upper quartiles.) In contrast, the distributions of design effects for Grades 4 and 11 appear quite similar--in exactly half of the 22 cases the median design effect for Grade 11 exceeds that for Grade 4.

The smaller design effects for Grade 8 indicate that the effects of the sample design on variance estimation is less for Grade 8 than for the other two grades. Since a major determinant of the sampling variability of a statistic is the degree of clustering in the sample, this would appear to be a surprising result. The major clustering in the sample is students within schools. Because the number of schools selected for assessment decreases by grade, with 661 schools selected at Grade 4, 478 at Grade 8 and 326 at Grade 11, the number of students selected within a school to respond to a given exercise increases by grade. On average, the number of students within a school responding to a given exercise is roughly three for Grade 4, five for Grade 8 and eight for Grade 11. All else being equal, this would imply that the design effects for Grade 4 would tend to be the smallest, those for Grade 11 the largest, and those for Grade 8 in-between. However, it is not only the sample cluster size that counts, but the heterogeneity of the full clusters, before subsampling, that influences the result. Grade 11 schools are larger and more heterogeneous, and this latter effect would reduce design effect for them. But, since these schools have larger sample clusters, this gain is more than offset. Grade 4 has smaller and more homogeneous schools, and higher correlations, but smaller samples per school. The observed phenomena is the combined effect.

We now turn to examining the distributions of design effects within subgroups of a given grade. The sampling variability of a subgroup, relative to the entire sample, depends on, among other things, how that subgroup is spread throughout the sample and what (weighted) proportion of the total sample is accounted for by the subgroup. For example, the white subgroup of the race/ethnicity variable is fairly evenly spread throughout the sample and accounts for more than 75 percent of the total sample (by weight) for each age. Consequently, the distribution of design effects for this subgroup closely resembles that of the total population.

The subgroups determined by sex and parental education are also fairly evenly spread across the sample. In these cases, however, a given subgroup is a smaller proportion of the total population. Consequently, any effects of cluster selection (students within schools) on the variance estimates should be reduced, relative to the total population, because there are fewer observations per cluster but roughly the same number of clusters. The result is a tendency for the design effects for these subgroups to be somewhat lower than those for the total.

On the other hand, the distributions of design effects by region, while roughly having the same median as the total sample, also have noticeably more variability (as measured by the inter-quartile range). This is because the partitioning of the entire sample into regions occurs at the PSU level and so a PSU is either entirely included or entirely excluded in the estimation of statistics at the regional level. Since the PSU is the level of aggregation used for variance estimation purposes, the estimated variances of regional level statistics are based on fewer degrees of freedom than are those of national level statistics. Consequently, the sampling variability of the regional level variance estimates must be larger.

Overall, although the distributions of design effects are different by subgroup, they are, perhaps, similar enough (at least within a grade) to select an overall composite value which is adequate for most purposes. Because Grade 8 appears to have lower design effects in general, it should probably be treated separately.

In choosing a composite design effect, some consideration must be made about the relative consequences of overestimating the variance as opposed to underestimating the variance. For example, adopting the position that an overestimate of the variance is as severe an error as an underestimate leads to using a composite which is near to the center of the distributions of the design effects. Possible composites of this type are the mean and median design effects. In the current data, the mean design effects are 1.5, 1.4 and 1.6 for Grades 4, 8 and 11, respectively. These are close to, but greater than, the median design effects: 1.4, 1.3 and 1.4.

Alternatively, one can adopt the position that it is a graver error to underestimate the variability of a statistic than to overestimate it. For example, Johnson and King (1986) examine estimation of variances using design effects (among other techniques) under assumption that the consequences of an underestimate are three times as severe as those of an overestimate of the same magnitude. Assuming that the distribution of design effects is roughly independent of the jackknife variance, so that the size of a design effect does not depend on the size of the variance, and adopting a loss function which is a weighted sum of absolute values of the deviations of predicted from actual with underestimates receiving three times the weight of overestimates, produces the upper quartile of the design effects as the composite value. The values of this composite, for Grades 4, 8 and 11, respectively, are 1.8, 1.6 and 1.8.

Table 14.2(1)

Distributions of Design Effects for Demographic Subgroups

Grade 4

Group	LoQ	Median	HiQ	Max	Mean	% <= 2.0	% <= 2.5
TOTAL	1.25	1.54	1.96	2.88	1.61	75.6	96.2
MALE	1.18	1.44	1.69	2.89	1.45	91.6	99.2
FEMALE	1.13	1.38	1.67	2.42	1.41	94.7	100.0
WHITE	1.23	1.55	1.85	3.42	1.60	79.4	95.4
BLACK	1.13	1.38	1.72	2.96	1.45	85.5	96.9
HISPANIC	1.05	1.46	1.86	5.34	1.55	80.2	93.9
OTHER	0.90	1.08	1.38	3.45	1.18	96.9	99.2
NE	1.12	1.53	2.20	4.66	1.72	67.2	84.7
SE	1.00	1.47	1.98	3.90	1.54	76.3	89.3
CENTRAL	1.05	1.62	2.27	4.50	1.71	64.1	83.2
WEST	0.93	1.30	1.98	5.20	1.55	75.6	84.0
< H.S.	1.06	1.30	1.68	3.01	1.39	87.0	99.2
GRAD HS	1.06	1.31	1.65	2.43	1.36	93.1	100.0
POST HS	1.11	1.38	1.69	3.03	1.42	90.1	98.5
UNKNOWN	1.13	1.37	1.55	2.57	1.38	95.4	99.2
RURAL	1.02	1.37	1.82	5.21	1.52	79.4	90.8
LOW MET	0.90	1.20	1.70	3.59	1.32	87.8	95.4
HI MET	1.17	1.56	1.95	4.14	1.63	75.6	87.8
BIG CITY	0.88	1.23	1.68	3.26	1.33	84.7	98.5
FRINGE	0.89	1.25	1.65	4.08	1.36	84.7	93.1
MED CITY	1.10	1.76	2.30	5.21	1.86	58.0	82.4
SMALL PL	1.05	1.40	1.70	3.14	1.43	84.7	96.9

Figure 14.2-1a

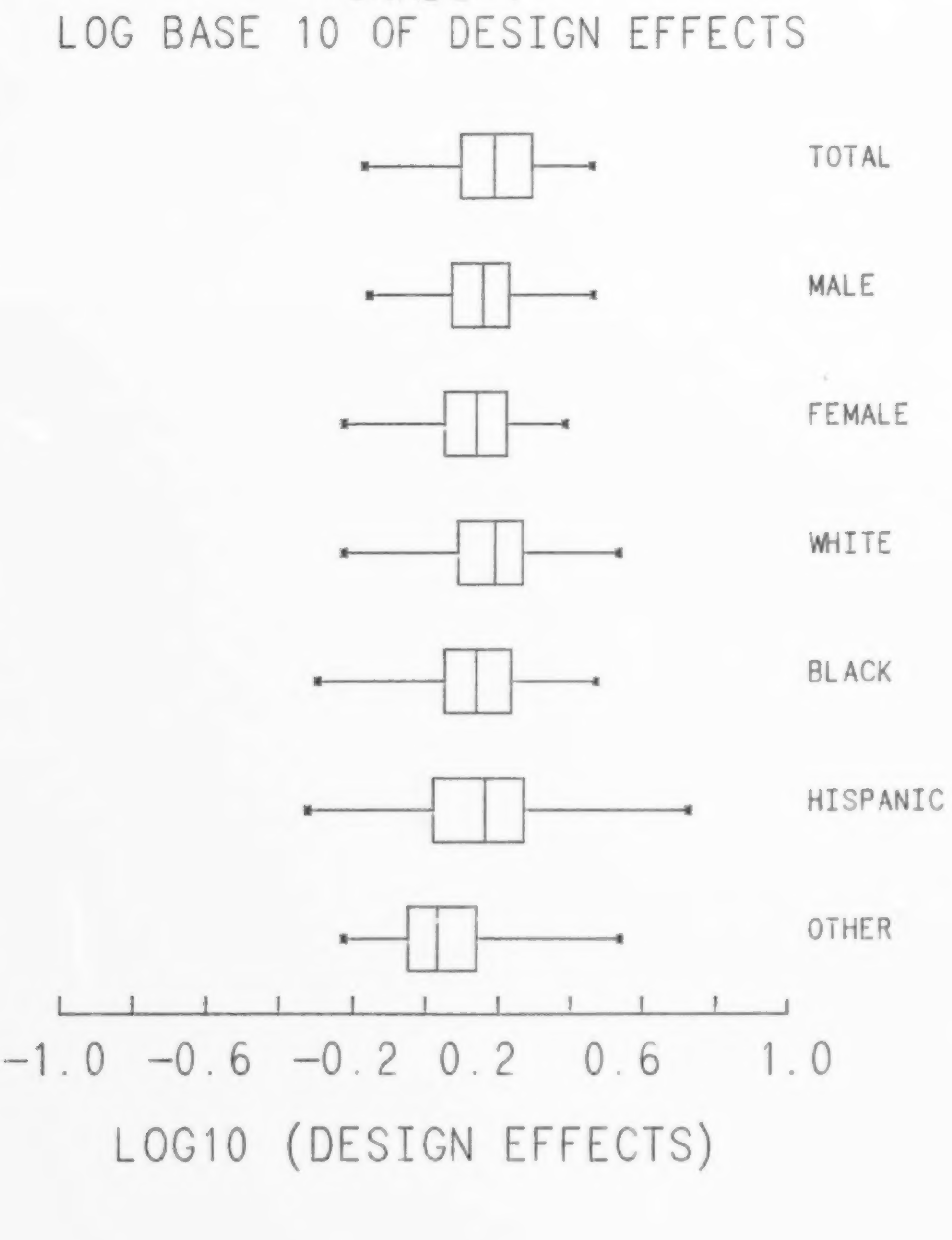


Figure 14.2-1b

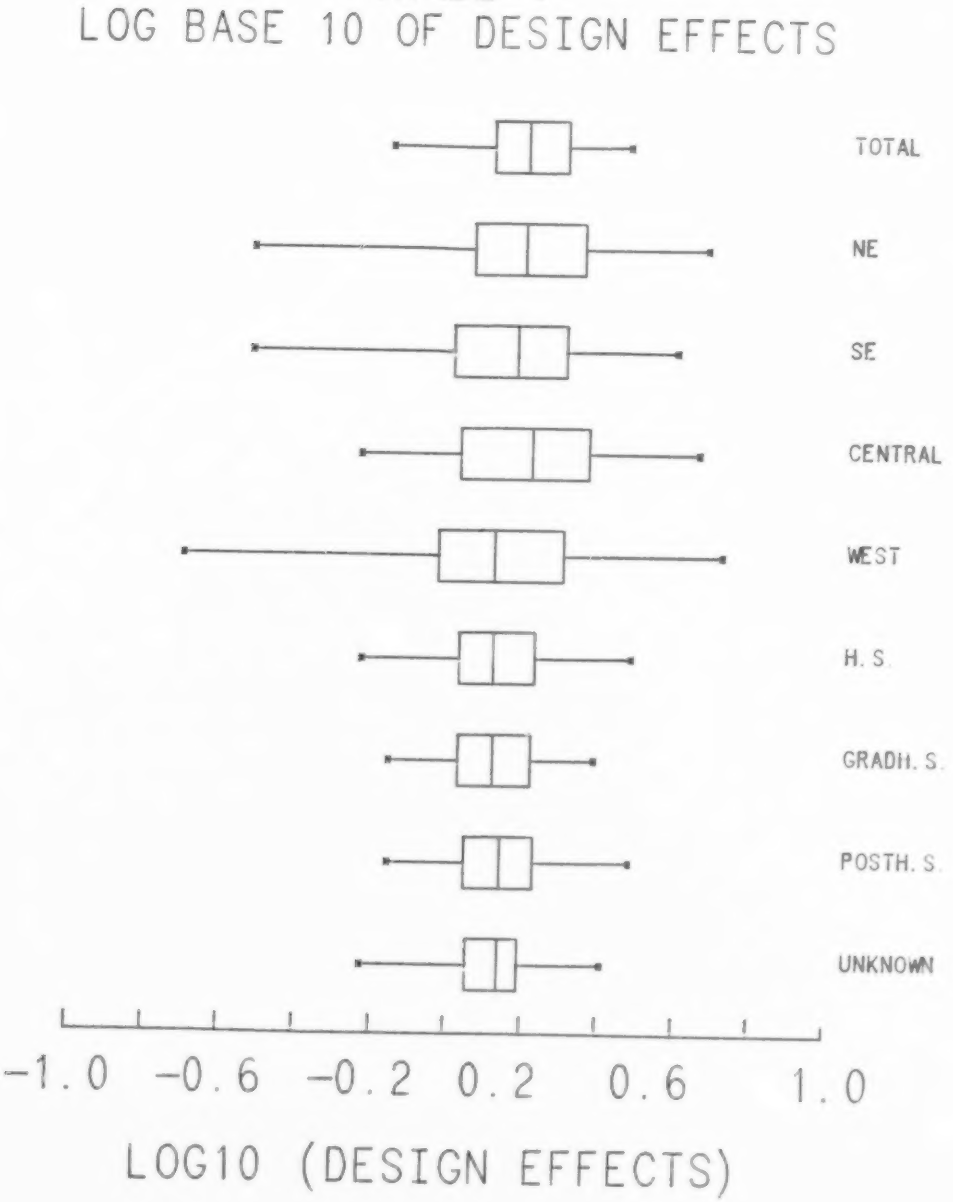


Figure 14.2-1c

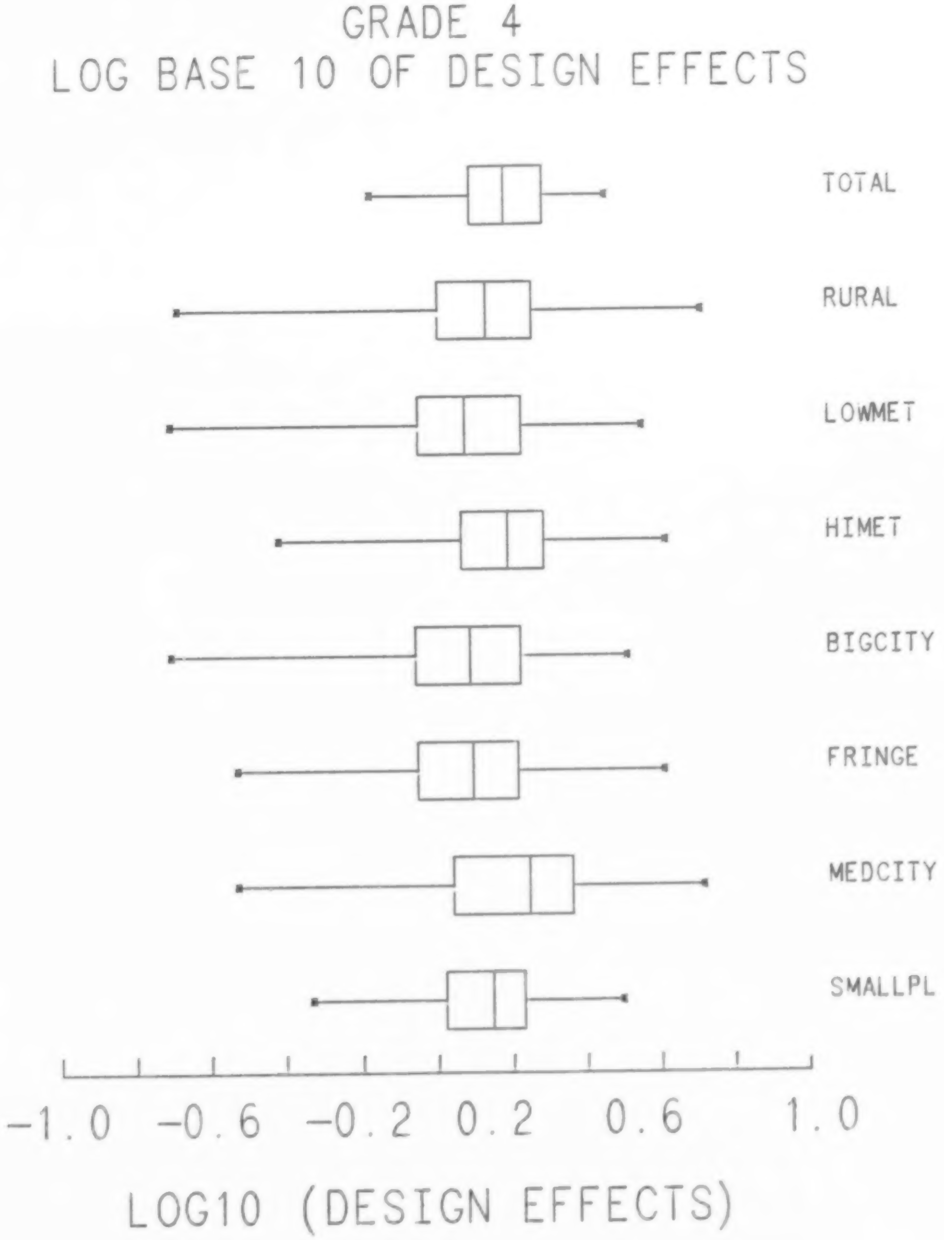


Table 14.2(2)

Distributions of Design Effects for Demographic Subgroups

Grade 8

Group	LoQ	Median	HiQ	Max	Mean	% <= 2.0	% <= 2.5
TOTAL	1.13	1.36	1.57	2.52	1.39	92.3	99.2
MALE	1.07	1.27	1.52	2.44	1.29	98.5	100.0
FEMALE	1.02	1.26	1.51	2.00	1.27	98.5	100.0
WHITE	1.09	1.32	1.53	2.79	1.32	94.6	99.2
BLACK	1.08	1.29	1.59	2.74	1.38	90.8	97.7
HISPANIC	0.97	1.33	1.87	4.68	1.54	80.0	86.9
OTHER	0.83	1.13	1.49	2.34	1.16	97.7	100.0
NE	0.83	1.23	1.78	3.57	1.38	78.5	89.2
SE	0.97	1.34	1.90	3.20	1.43	80.0	91.5
CENTRAL	0.87	1.33	1.83	4.03	1.47	80.0	87.7
WEST	0.86	1.23	1.64	3.61	1.30	88.5	96.2
< H.S.	0.98	1.23	1.51	2.92	1.27	93.1	99.2
GRAD HS	1.07	1.27	1.46	2.18	1.29	96.2	100.0
POST HS	0.96	1.17	1.45	2.70	1.21	98.5	99.2
UNKNOWN	0.93	1.16	1.38	2.43	1.20	98.5	100.0
RURAL	0.86	1.39	1.92	4.14	1.45	77.7	92.3
LOW MET	0.89	1.37	1.90	4.14	1.49	76.9	88.5
HI MET	0.98	1.39	2.05	5.03	1.60	73.8	86.2
BIG CITY	0.75	1.12	1.76	5.12	1.35	80.8	87.7
FRINGE	0.60	0.94	1.36	3.95	1.10	89.2	95.4
MED CITY	1.13	1.64	2.18	4.89	1.76	65.4	81.5
SMALL PL	0.90	1.26	1.61	2.85	1.31	87.7	97.7

Figure 14.2-2a

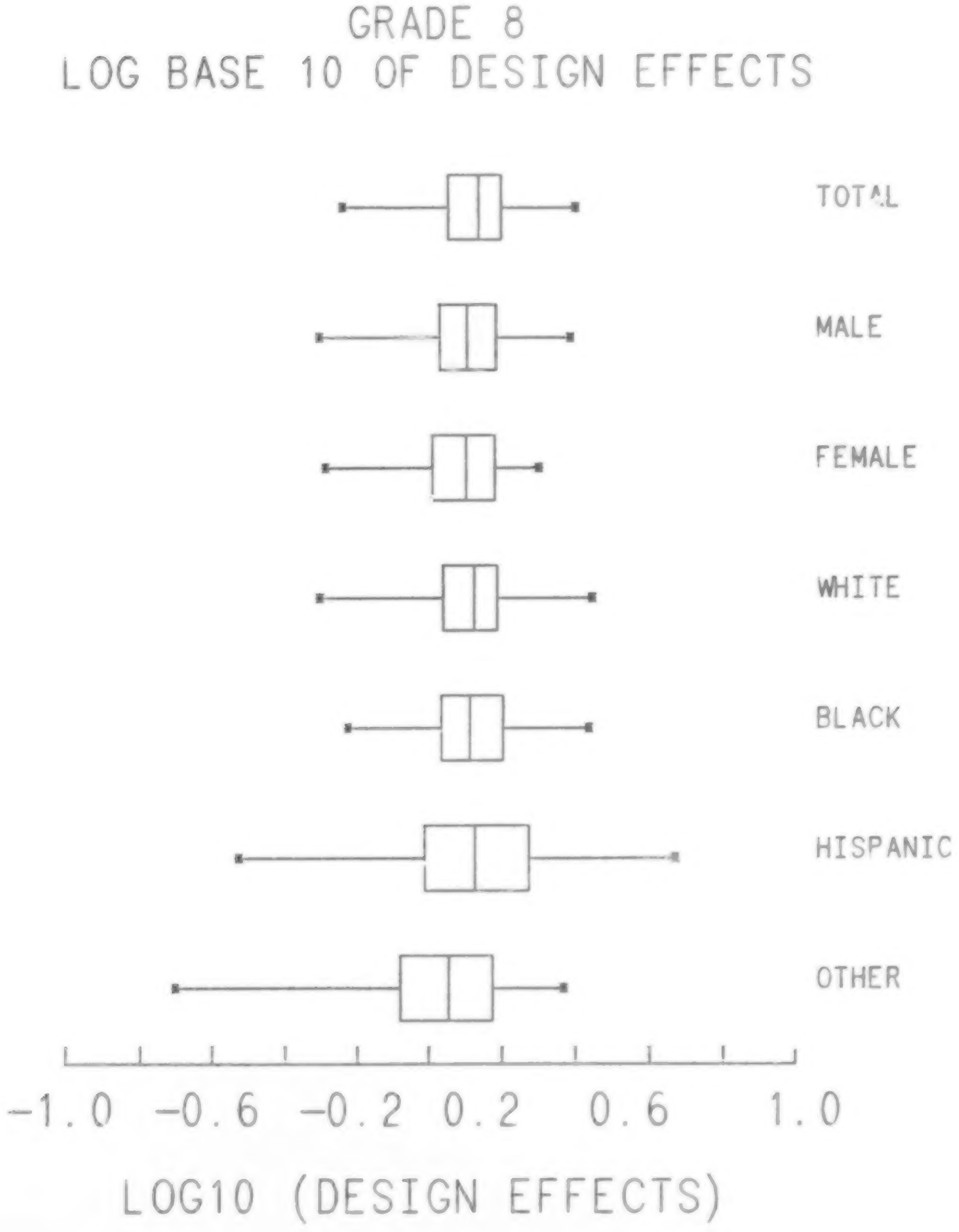


Figure 14.2-2b

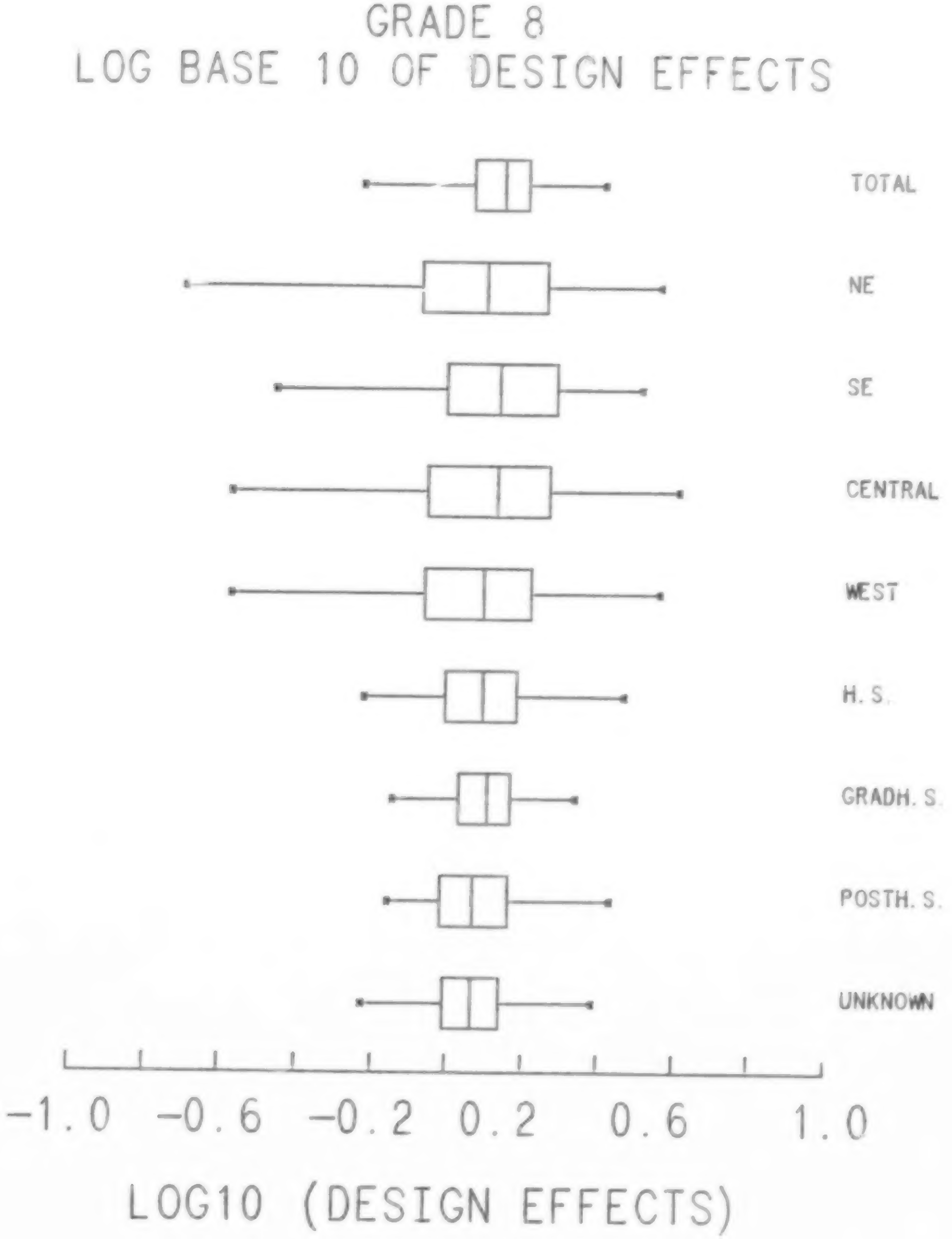


Figure 14.2-2c

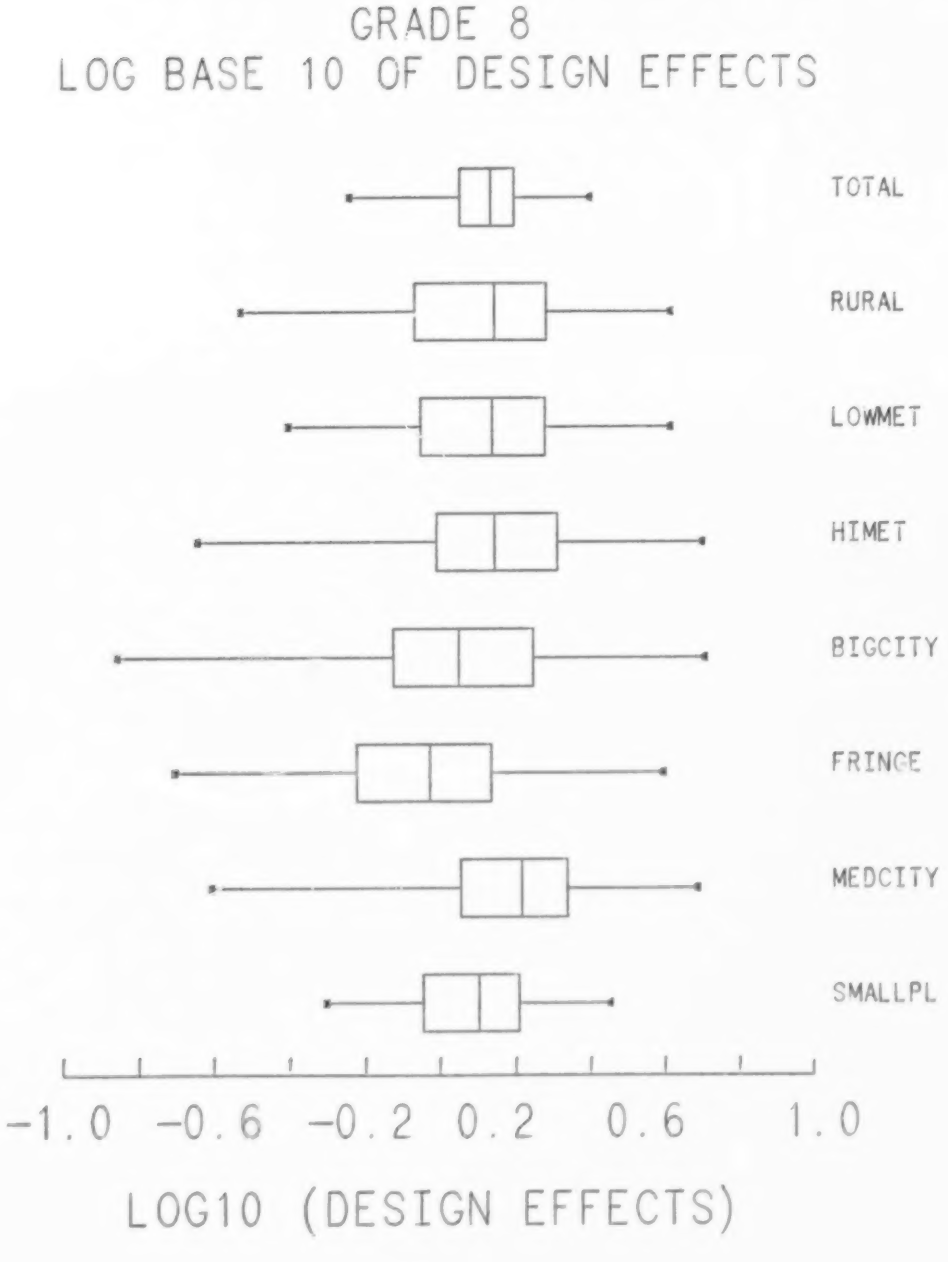


Table 14.2(3)

Distributions of Design Effects for Demographic Subgroups

Grade 11

Group	LoQ	Median	HiQ	Max	Mean	% <= 2.0	% <= 2.5
TOTAL	1.23	1.55	1.95	2.84	1.63	75.9	92.2
MALE	1.12	1.32	1.65	2.80	1.42	90.5	98.3
FEMALE	1.21	1.46	1.75	3.65	1.55	85.3	92.2
WHITE	1.24	1.55	1.83	3.24	1.57	84.5	95.7
BLACK	1.13	1.54	2.04	3.60	1.70	71.6	83.6
HISPANIC	0.97	1.43	2.02	3.40	1.52	74.1	93.1
OTHER	0.94	1.16	1.40	2.60	1.21	92.2	99.1
NE	1.35	1.89	2.65	5.51	2.10	53.4	71.6
SE	0.96	1.44	2.02	5.42	1.57	74.1	90.5
CENTRAL	0.97	1.48	2.02	4.32	1.59	75.0	87.1
WEST	0.80	1.28	1.73	4.24	1.37	81.0	90.5
< H.S.	1.12	1.35	1.63	2.42	1.41	88.8	100.0
GRAD HS	1.07	1.30	1.59	2.60	1.37	91.4	99.1
POST HS	1.24	1.46	1.73	2.76	1.52	86.2	95.7
UNKNOWN	0.96	1.17	1.47	2.53	1.22	95.7	99.1
RURAL	1.05	1.42	1.93	6.21	1.59	77.6	92.2
LOW MET	1.12	1.57	2.37	5.31	1.85	65.5	77.6
HI MET	0.99	1.60	2.29	6.54	1.82	67.2	79.3
BIG CITY	0.78	1.09	1.57	3.08	1.25	87.1	94.0
FRINGE	0.77	1.01	1.38	3.40	1.12	93.1	97.4
MED CITY	0.79	1.24	1.78	4.29	1.39	78.4	90.5
SMALL PL	1.03	1.36	1.77	2.98	1.45	81.9	93.1

Figure 14.2-3a

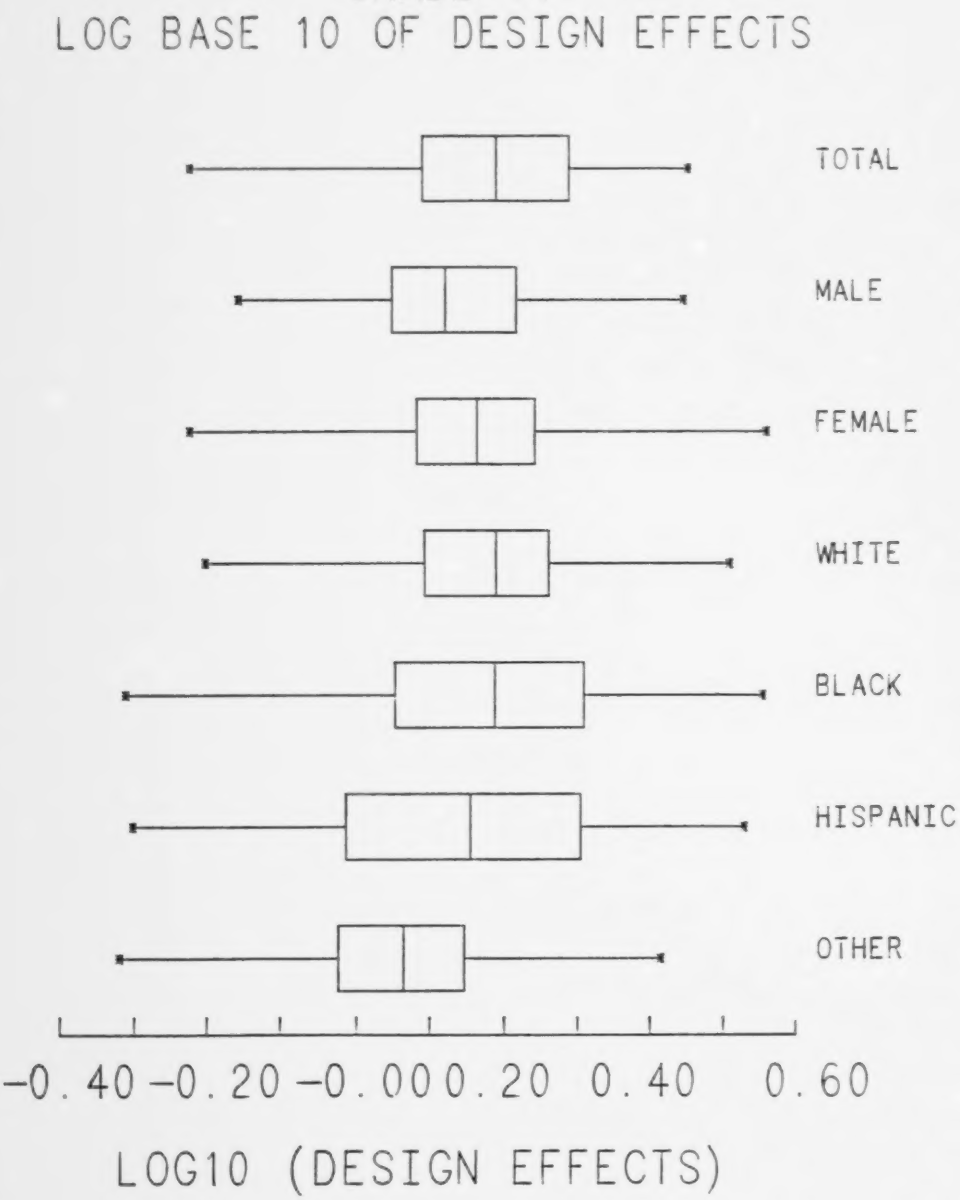


Figure 14.2-3b

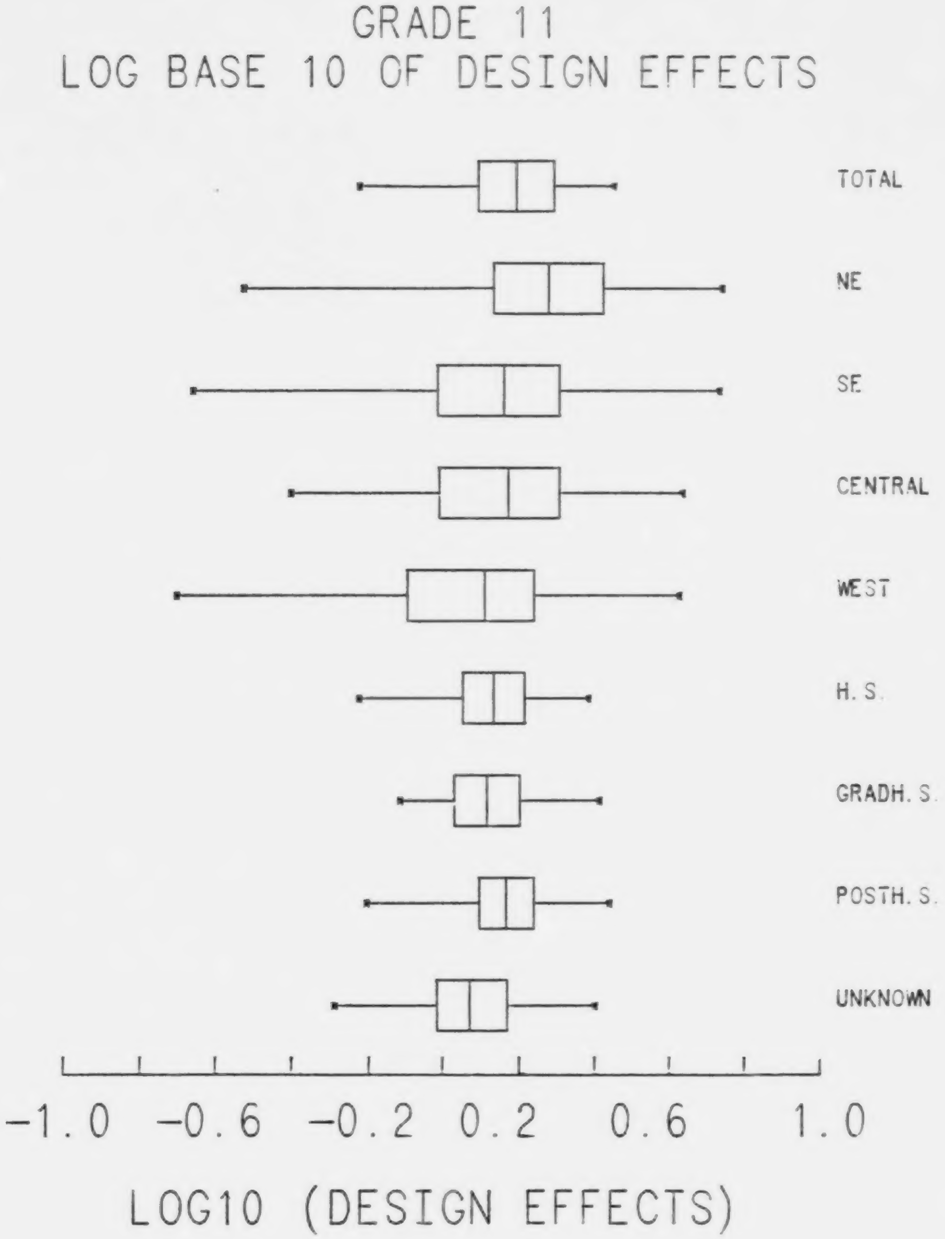
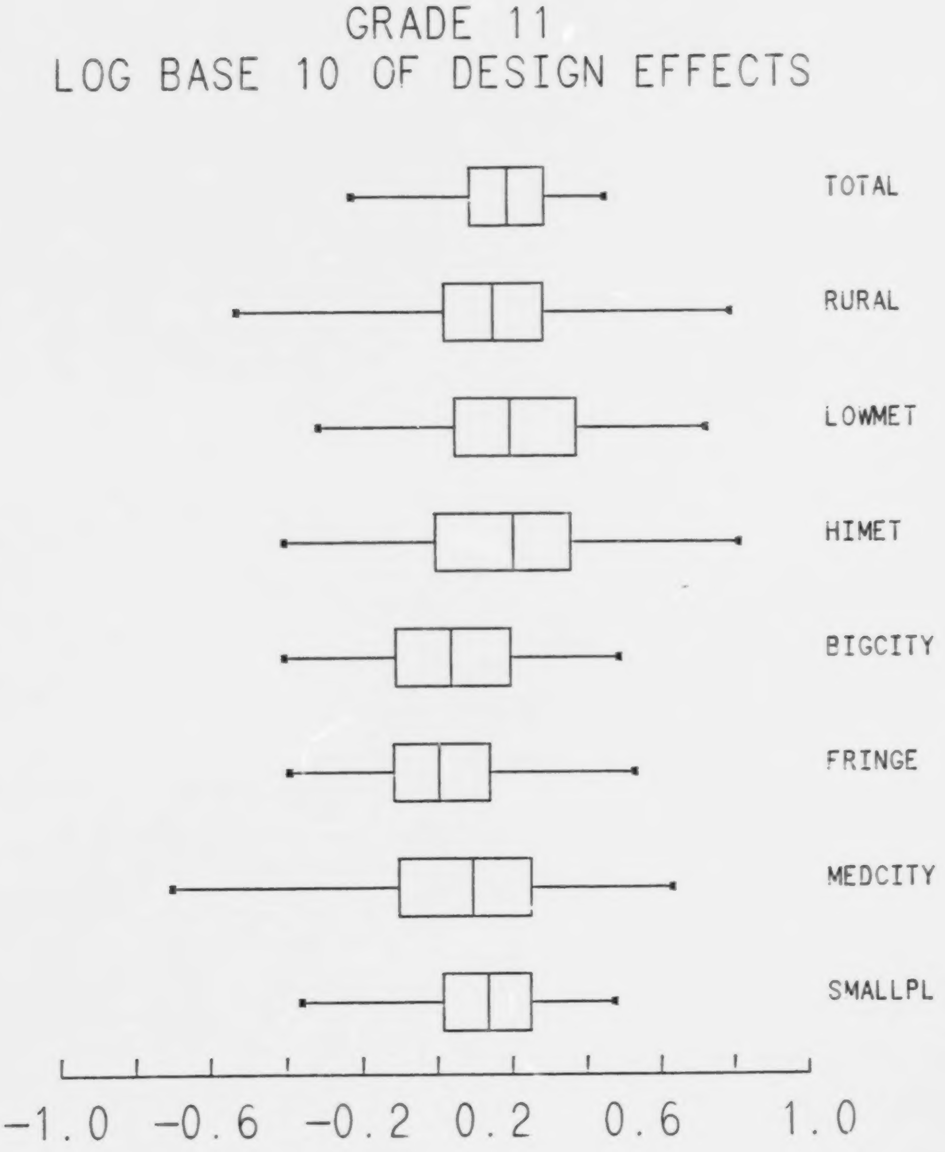


Figure 14.2-3c



IMPLEMENTING THE NEW DESIGN:
THE NAEP 1983-84 TECHNICAL REPORT

PART III

Chapter 15

ESTIMATES OF THE READING AND WRITING PROFICIENCY OF AMERICAN STUDENTS

Albert E. Beaton
David S. Freund
Bruce A. Kaplan

Educational Testing Service

This part of the technical report presents estimates of the reading and writing proficiency of students in American schools. The first part of this report described how the students were selected, how they were assessed, and how their responses moved from assessment sessions to a carefully constructed data base, ready for analysis. The second part described the methods of data analysis, including scaling and parameter estimation. In this third part, estimates of how students are performing in school, and estimates of the sampling error, are presented.

This is a technical report and is not intended to be interpretive. Estimates are presented, but no attempt is made to explain why the students behaved in the way that they did. Interpretive results are presented in NAEP reports such as The Reading Report Card: Progress Toward Excellence in Our Schools (1985), and Writing: Trends Across the Decade, 1974-84 (Applebee, Langer, & Mullis, 1986a). We will leave it to experts in the educational process to hypothesize why the results occurred. We have made the public-use data tapes (Barone, Norris, & Rogers, 1986) available for those who wish to estimate other functions of student performance from the NAEP data or to search for possible explanations for the student performance that is reported here.

Clearly, neither this report, nor any report, could present all of the population estimates that are made possible by the NAEP database. The analysis of the 1983-84 NAEP data has resulted in the production of many thousands of tables containing estimates of the proficiency of students, and various subgroups of students, in American schools. These tables have been bound in books called almanacs; the contents of 24 such almanacs are described in Chapter 13.4. We have selected a few of the most basic tables for presentation here. In addition, some tables that are not included in almanacs are presented.

The technical details of the estimation process which underlies these tables are covered in the previous parts of this report and not repeated here. For a detailed description of how to read and use the tables selected from the almanacs the reader should refer to Chapter 13.4. The computational procedures for other tables will be noted as needed.

15.1 Population Estimates

The NAEP Year 15 data includes a number of different samples from which population estimates can be made, and Westat, Inc. has developed an appropriate set of sampling weights for the students in each sample. All estimates of population parameters use these sampling weights.

Table 15(1) shows the sizes of the various samples and the sums of their sampling weights by grade/age combination. The sums of the weights for the spiral samples, which are by far the largest, estimate the numbers of students who are in each grade/age combination and who would be assessable. The sums of the weights of the excluded students estimate the numbers of students in each grade/age combination who, in their schools' judgment, would not be assessable. The sums of the estimates for the spiral sample and for the excluded sample are estimates of the total number of in-school students in the grade/age combination.

The four tape samples are defined by age only, and each sum of weights is an estimate of the number of age-eligible students in an age category who, in their school's judgment, would be assessable. These weight sums can be added to the sum of weights of the age-eligible excluded students to make an estimate of the number of students in an age category. The differences in the estimates from the four tape samples are due to sampling error.

In most cases, the number of students in a grade/age combination is not of interest; a researcher will be interested in estimating the number of students at either a grade or an age. An estimate of the number of students at an age level can be made by summing the weights of only the age-eligible students, and an estimate of the number of students in a grade by summing the weights of grade-eligible students.

Table 15(2) shows how many students at each grade level are at, in, or above the modal age for that grade, and how many at each age level are at, in, or above the modal grade for that age. These figures were computed from the spiral sample only. Along with the counts from this sample, the sum of the weights (Weighted N) for each category is presented, and these sums are estimates of the numbers of students in these categories in the population. The standard errors of these estimates and coefficients of variation are also given.

Tables 15(3), 15(4), and 15(5) present estimates of the number of students in various subpopulations who could have been assessed for each of the grade/age combinations. These estimates were made from the spiral sample. Separate estimates are shown for the different sexes, racial/ethnic groupings, regions of the country, levels of parental education, and sizes and types of community. Estimates are made separately for age-eligible and grade-eligible students as well as for students who are eligible by both age and grade and those who are eligible by either age or grade. The actual numbers of students used in making these estimates are shown in Tables 2(10), 2(11) and 2(12) of Chapter 2.

Tables 15(6), 15(7), and 15(8) present estimates of the number of students in various subpopulations who, in their schools' judgment, were unassessable. Separate estimates are also shown for the different sexes, racial/ethnic groupings, regions of the country, classes of parental education, and sizes and types of community. Also, estimates are made separately for age-eligible and grade-eligible students as well as for students who are eligible by both age and grade and those who are eligible by either age or grade. The sum of the estimates from these tables and the corresponding estimates from the previous tables are estimates of the total number of students at an age or grade level, whether they were assessable or not. The actual numbers of students used in making these estimates are shown in Tables 2(13), 2(14) and 2(15) of Chapter 2.

Tables 15(9), 15(10), and 15(11) show estimates of the numbers of assessable students at different age levels. There are four estimates, one made from each of the four tape samples. Separate estimates are also shown for the different sexes, racial/ethnic groupings, regions of the country, classes of parental education, and sizes and types of community. The average of the four estimates is also an estimator of the population size. These samples cannot be used to estimate grade populations. The actual numbers of students used in making these estimates are shown in Tables 2(16), 2(17) and 2(18) of Chapter 2.

Tables 15(12), 15(13), and 15(14) were produced from the spiral sample and are included as background for the tables that follow. These tables show the actual numbers of students for whom plausible values were computed and population estimates of the numbers who would have had plausible values if an educational census of the entire country were done using this NAEP design. The design for NAEP called for some students to be assigned reading exercises, some to be assigned writing exercises, and some assigned both. Plausible values for reading were computed for only those students who were assigned reading exercises that were used in the scaling process, and similarly for writing plausible values. This table reinforces the fact that plausible values for writing were computed for only those students who were eligible in the grade samples, not the age samples.

The following tables present reading and writing proficiency estimates.

- Table 15(15) displays reading proficiency estimates for fourth graders.
- Table 15(16) displays reading proficiency estimates for eighth graders.
- Table 15(17) displays reading proficiency estimates for eleventh graders.

- Table 15(18) displays writing proficiency estimates for fourth graders.
- Table 15(19) displays writing proficiency estimates for eighth graders.
- Table 15(20) displays writing proficiency estimates for eleventh graders.

Population estimates are presented for students of specified grade levels only (i.e., grades 4, 8, and 11). These tables also contain separate proficiency estimates for the same selected population subgroups as in the preceding tables but are restricted to the students in the specified grades. In particular, breakdowns by age within grade are given. Since the students of a given grade can be of any age, it is important to note that some of these age estimates, which are conditional on grade placement, are made from small samples and none of these estimates are appropriate for estimating the proficiency of the entire population at an age level.

For all assessable students in a grade, and for each subgroup, these tables contain the actual sample sizes used in computing the estimates and the sum of the weights (Weighted N) for those samples. In these tables, the Weighted N will be of little interest, but the coefficient of variation, which is in parentheses next to the Weighted N, is a measure of the variability of the estimates of the standard errors and thus of importance in judging the adequacy of the population estimates. Large coefficients of variation (exceeding 20 percent) are emphasized with an exclamation point (!) in the tables.

Next, for each subgroup and the total, the tables contain estimates of mean values, standard deviations[1], and 10th, 25th, 50th (median), 75th, and 90th percentiles. Each of these statistics is followed by an estimate of its standard error, which is in parentheses.

Tables 15(21) through 15(69) report either estimated average values or percents below anchor points on the reading or writing scales. These tables, which have been selected from several proficiency almanacs, are based on the spiral grade-eligible sample only. A detailed discussion of how these tables were constructed and how to read and use these tables is presented in Chapter 13.

---

[1] These standard deviations were computed as follows: (1) for each of five plausible values, a weighted analogue of the conventional sample variance was computed; (2) the square roots of each of these estimates was taken; and (3) the average of the five values in (2) was obtained. The weighted variances computed in (1) do not take into account the sample design (see Tepping & Hansen, 1984). However, attempts to implement an adjustment did not lead to satisfactory results (Zwick, 1985).

There are six sets of eight tables each:

- Tables 15(21) through 15(28) contain reading proficiency values for fourth graders.
- Tables 15(29) through 15(36) contain reading proficiency values for eighth graders.
- Tables 15(37) through 15(44) contain reading proficiency values for eleventh graders.
- Tables 15(45) through 15(52) contain writing proficiency values for fourth graders.
- Tables 15(53) through 15(60) contain writing proficiency values for eighth graders.
- Tables 15(61) through 15(68) contain writing proficiency values for eleventh graders.

These tables contain separate estimates for the different sexes, racial/ethnic groups, levels of parental education, and ages of the students who were in those grades[2]. These groups are sometimes referred to as reporting subgroups and are defined in Chapter 12. In some tables, some age groups were so small that estimates were not reported.

Each set of tables contains:

(1) Estimates of the average performance of the total grade and for each reporting group.

(2) Estimates of the average performance of the males and females in each reporting group.

(3) Estimates of the average performance by racial/ethnic grouping (Whites, Blacks, Hispanics, American Indians, Asians, and Unclassified) in each reporting group.

---

[2] The estimates of subgroup reading proficiency reported in Tables 15(21) through 15(44) are occasionally different from the corresponding estimates reported in Tables 15(15) through 15(17), although the difference is always trivial. The discrepancy occurs because the estimates of performance in Tables 15(21) through 15(44) are based on a single set of plausible values, while the estimates in Tables 15(15) through 15(17) are based on five sets of plausible values. While both sets of estimates have equal expectations, the estimates based on the average of five separate estimates are less variable. Both sets of estimates used the same estimates of sampling variability, which includes a component of variability of estimate across sets of plausible values.

(4) Estimates of the average performance of the students from various regions of the country (Northeast, Southeast, Central, and Western) in each reporting group.

(5) Estimates of the average performance of students at various age levels (conditional on grade) within each reporting group.

(6) Estimates of the average performance of students from different sizes and types of communities (Rural, Disadvantaged Urban, Advantaged Urban, Other Big City, Fringe of Big City, Medium Cities, and Small Cities within each reporting group) within each reporting group.

(7) Estimates of the average performance of students by self-reported parents' education (Did Not Graduate High School, Did Graduate High School, Had Some Post-High School Education, or Unknown) within each reporting group.

(8) Estimates of the percents above selected anchor points for each reporting group.

Along with each estimated proficiency value is the estimated proportion of the population comprising that category. For example, the TOTAL line of Table 15(22) shows, among other things, that the estimated average reading proficiency score of male fourth-graders is 215.1 and that males are estimated to constitute 49.8 percent of the fourth-grade population. Each estimated average value and percent is accompanied by its standard error.

These tables contain some redundancy; for example, sex is used both as a reporting variable and to classify the students within reporting variables. In these cases, the logically impossible categories in the tables are replaced by *****.

Table 15(1)

Number of Students by Grade/Age Combination and by Type of Assessment

	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
	Count	Sum of Weights	Count	Sum of Weights	Count	Sum of Weights
SPIRAL	26087	3971749	28405	4363428	28861	4045041
TAPE 1	1403	3122045	1310	3310355	1539	3048025
TAPE 2	1356	3005769	1276	3348263	1540	3045283
TAPE 3	1389	3087985	1283	3338947	1596	2978520
TAPE 4	1344	3100713	1289	3340216	1534	3026687
EXCLUDED	1416	171436	1448	179054	1361	115162

Table 15(2)

Number of Spiral-Assessed Students by Grade/Age

Grade 4/Age 9

	AGE			
	< 9	= 9	> 9	TOTAL
GRADE < 4				
UNWEIGHTED N	0	5917	0	5917
WEIGHTED N	0	761524	0	761524
STANDARD ERROR	-	8099	-	8099
COEFF. OF VAR.	-	1.06	-	1.06
GRADE = 4				
UNWEIGHTED N	158	12953	6984	20095
WEIGHTED N	21504	2295588	883198	3200290
STANDARD ERROR	2395	9067	7089	14047
COEFF. OF VAR.	11.14	0.39	0.80	0.44
GRADE > 4				
UNWEIGHTED N	0	75	0	75
WEIGHTED N	0	9936	0	9936
STANDARD ERROR	-	2163	-	2163
COEFF. OF VAR.	-	21.77	-	21.77
GRADE TOTAL				
UNWEIGHTED N	158	18945	6984	26087
WEIGHTED N	21504	3067047	883198	3971749
STANDARD ERROR	2395	13693	7089	18935
COEFF. OF VAR.	11.14	0.45	0.80	0.48

Table 15(2)
(continued)

Number of Spiral-Assessed Students by Grade/Age

Grade 8/Age 13

	AGE			
	< 13	= 13	> 13	TOTAL
GRADE < 8				
UNWEIGHTED N	0	6495	0	6495
WEIGHTED N	0	1034711	0	1034711
STANDARD ERROR	-	6087	-	6087
COEFF. OF VAR.	-	0.59	-	0.59
GRADE = 8				
UNWEIGHTED N	184	14515	7151	21850
WEIGHTED N	25662	2269841	1018446	3313949
STANDARD ERROR	4070	4025	6860	7824
COEFF. OF VAR.	15.86	0.18	0.67	0.24
GRADE > 8				
UNWEIGHTED N	0	60	0	60
WEIGHTED N	0	14769	0	14769
STANDARD ERROR	-	4144	-	4144
COEFF. OF VAR.	-	28.06	-	28.06
GRADE TOTAL				
UNWEIGHTED N	184	21070	7151	28405
WEIGHTED N	25662	3319320	1018446	4363428
STANDARD ERROR	4070	8088	6860	11281
COEFF. OF VAR.	15.86	0.24	0.67	0.26

Table 15(2)
(continued)

Number of Spiral-Assessed Students by Grade/Age

Grade 11/Age 17

	AGE			
	< 17	= 17	> 17	TOTAL
GRADE < 11				
UNWEIGHTED N	0	4129	0	4129
WEIGHTED N	0	671683	0	671683
STANDARD ERROR	-	21099	-	21099
COEFF. OF VAR.	-	3.14	-	3.14
GRADE = 11				
UNWEIGHTED N	2386	16787	3692	22865
WEIGHTED N	399289	2037738	635595	3072622
STANDARD ERROR	22106	3439	21691	6816
COEFF. OF VAR.	5.54	0.17	3.41	0.22
GRADE > 11				
UNWEIGHTED N	0	1867	0	1867
WEIGHTED N	0	300736	0	300736
STANDARD ERROR	-	20072	-	20072
COEFF. OF VAR.	-	6.67	-	6.67
GRADE TOTAL				
UNWEIGHTED N	2386	22783	3692	28861
WEIGHTED N	399289	3010157	635595	4045041
STANDARD ERROR	22106	8221	21691	11104
COEFF. OF VAR.	5.54	0.27	3.41	0.27

Table 15(3)

Estimated Total Number of Students in the Population Eligible for Assessment

Grade 4/Age 9

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3067047	3200290	2295588	3971749
SEX:				
MALE	1507250	1594774	1073745	2028280
FEMALE	1559795	1605515	1221842	1943468
RACE:				
WHITE	2173213	2260759	1664721	2769252
BLACK	442833	486025	317635	611224
HISPANIC	359832	362177	247662	474347
OTHER	91168	91328	65570	116926
REGION:				
NORTHEAST	675001	714539	513833	875707
SOUTHEAST	727163	762456	542387	947232
CENTRAL	822401	851344	618875	1054869
WEST	842481	871951	620492	1093939
PARENTS ED:				
LESS THAN HIGH SCHOOL	165907	190429	115449	240887
HIGH SCHOOL	591450	641628	447221	785857
GREATER THAN HIGH SCHOOL	1137477	1211729	906596	1442611
UNKNOWN	1172211	1156503	826322	1502393
SIZE AND TYPE OF COMMUNITY:				
RURAL	197666	206921	142139	262448
DISADVANTAGED URBAN	373042	401687	280907	493821
ADVANTAGED URBAN	427890	454718	349044	533565
BIG CITY	242164	239410	179708	301866
FRINGE	340643	353975	257068	437549
MEDIUM	501314	515287	363553	653048
SMALL	984328	1028292	723168	1289451

Table 15(4)

Estimated Total Number of Students in the Population Eligible for Spiral Assessment

Grade 8/Age 13

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3319320	3313949	2269841	4363428
SEX:				
MALE	1672496	1664184	1067848	2268831
FEMALE	1646719	1649513	1201887	2094346
RACE:				
WHITE	2462505	2447689	1758484	3151710
BLACK	470115	483370	281652	671833
HISPANIC	291061	290652	161122	420591
OTHER	95639	92238	68583	119294
REGION:				
NORTHEAST	760547	757155	522568	995134
SOUTHEAST	766165	776671	523458	1019378
CENTRAL	890588	878077	611431	1157234
WEST	902019	902046	612383	1191681
PARENTS ED:				
LESS THAN HIGH SCHOOL	286535	312820	164915	434441
HIGH SCHOOLS	1164547	1170291	792415	1542423
GREATER THAN HIGH SCHOOL	1512740	1512803	1125597	1899946
UNKNOWN	355497	318034	186913	486618
SIZE AND TYPE OF COMMUNITY:				
RURAL	174221	176480	117124	233576
DISADVANTAGED URBAN	297675	294289	182362	409602
ADVANTAGED URBAN	356948	352477	275203	434222
BIG CITY	359003	346473	244980	460496
FRINGE	543237	549813	381908	711142
MEDIUM	484395	499186	323148	660433
SMALL	1103841	1095230	745115	1453957

Table 15(5)

Estimated Total Number of Students
in the Population Eligible for Spiral Assessment

Grade 11/Age 17

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3010157	3072622	2037738	4045041
SEX:				
MALE	1532749	1548465	979337	2101878
FEMALE	1477134	1524155	1058401	1942889
RACE:				
WHITE	2259880	2284441	1602266	2942055
BLACK	426035	458544	248863	635716
HISPANIC	242188	245739	135384	352543
OTHER	82053	83897	51224	114726
REGION:				
NORTHEAST	734255	754812	482051	1007016
SOUTHEAST	669005	675109	418403	925711
CENTRAL	816774	840587	597192	1060168
WEST	790123	802113	540092	1052144
PARENTS ED:				
LESS THAN HIGH SCHOOL	351212	350297	192273	509236
HIGH SCHOOL	1041252	1031742	688507	1384486
GREATER THAN HIGH SCHOOL	1491560	1567083	1090224	1969219
UNKNOWN	126133	122699	66733	182098
SIZE AND TYPE OF COMMUNITY:				
RURAL	153730	166812	107855	212686
DISADVANTAGED URBAN	308175	321553	163095	466633
ADVANTAGED URBAN	482440	506711	345115	644037
BIG CITY	266997	270523	169867	367652
FRINGE	327360	316221	225178	418403
MEDIUM	493442	513958	355848	651551
SMALL	978013	976844	670779	1284078

Table 15(6)

Estimated Total Number of Students in the Population Eligible for Spiral Assessment Who Would Be Deemed Unassessable by Their Schools

Grade 4/Age 9

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	107871	105233	41668	171436
SEX:				
MALE	68743	66881	25622	110002
FEMALE	38971	37895	15889	60977
RACE:				
WHITE	50802	47171	17982	79991
BLACK	13227	14178	5343	22062
HISPANIC	32288	32882	13966	51204
OTHER	11554	11002	4377	18178
REGION:				
NORTHEAST	23083	21263	10793	33553
SOUTHEAST	24597	20449	7366	37680
CENTRAL	21528	22964	8042	36451
WEST	38663	40557	15468	63752
SIZE AND TYPE OF COMMUNITY:				
RURAL	4784	8270	1301	11752
DISADVANTAGED URBAN	21558	22815	8568	35805
ADVANTAGED URBAN	11483	10004	6102	15385
BIG CITY	8818	5538	3531	10826
FRINGE	15320	16117	8775	22662
MEDIUM	18129	16158	4383	29905
SMALL	27778	26331	9009	45100

Table 15(7)

Estimated Total Number of Students in the Population Eligible for Spiral Assessment Who Would Be Deemed Unassessable by Their Schools

Grade 8/Age 13

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	101041	116028	38014	179054
SEX:				
MALE	65191	72953	23634	114510
FEMALE	35202	42817	14310	63709
RACE:				
WHITE	52073	64039	20360	95752
BLACK	19199	20180	6138	33241
HISPANIC	18922	20775	7312	32385
OTHER	10847	11034	4206	17676
REGION:				
NORTHEAST	20033	22252	8419	33866
SOUTHEAST	21251	27801	6604	42449
CENTRAL	31571	37884	12239	57216
WEST	28185	28091	10752	45523
SIZE AND TYPE OF COMMUNITY:				
RURAL	3273	6830	1448	8655
DISADVANTAGED URBAN	18243	20111	8613	29742
ADVANTAGED URBAN	6287	6721	1686	11321
BIG CITY	13994	14529	5683	22840
FRINGE	12514	11733	5026	19221
MEDIUM	14764	19358	3394	30728
SMALL	31966	36746	12165	56548

Table 15(8)

Estimated Total Number of Students in the Population Eligible for Spiral Assessment Who Would Be Deemed Unassessable by Their Schools

Grade 11/Age 17

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE & GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	74451	65829	25119	115162
SEX:				
MALE	47857	41745	16471	73131
FEMALE	26513	23983	8647	41849
RACE:				
WHITE	30787	29275	11238	48824
BLACK	15712	11073	3752	23034
HISPANIC	17957	15626	7159	26424
OTHER	9996	9855	2970	16881
REGION:				
NORTHEAST	11246	10482	3689	18039
SOUTHEAST	18700	14775	5540	27935
CENTRAL	19661	19786	6157	33290
WEST	24844	20787	9732	35899
SIZE AND TYPE OF COMMUNITY:				
RURAL	3158	3107	1201	5064
DISADVANTAGED URBAN	18443	13928	6433	25938
ADVANTAGED URBAN	5138	5526	2267	8396
BIG CITY	8434	6976	2521	12889
FRINGE	5969	6398	2246	10121
MEDIUM	13388	12010	4096	21302
SMALL	19921	17885	6354	31452

Table 15(9)

Estimated Total Number of Students Who Are Eligible for Assessment by Tape Sample

Age 9

	TAPE 1	TAPE 2	TAPE 3	TAPE 4
TOTAL	3122045	3005769	3087985	3100713
SEX:				
MALE	1541569	1546943	1547829	1502897
FEMALE	1580474	1458825	1540155	1597815
RACE:				
WHITE	2232047	2156800	2201732	2214269
BLACK	446612	434192	454103	451488
HISPANIC	331819	311022	323277	323472
OTHER	111566	103755	108873	111484
REGION:				
NORTHEAST	672037	660320	678638	648949
SOUTHEAST	778707	674785	774924	871961
CENTRAL	787159	900433	841201	768386
WEST	884142	770231	793221	811416
PARENTS ED:				
LESS THAN HIGH SCHOOL	186493	145728	178006	229364
HIGH SCHOOL	541633	613470	607155	645405
GREATER THAN HIGH SCHOOL	1279406	1090678	1179219	1097028
UNKNOWN	1114511	1155892	1123604	1128914
SIZE AND TYPE OF COMMUNITY:				
RURAL	319513	241335	237176	115983
DISADVANTAGED URBAN	317826	388177	508894	447247
ADVANTAGED URBAN	532838	464262	423289	250761
BIG CITY	270103	117433	366492	319272
FRINGE	266304	355329	85568	243818
MEDIUM	603469	241238	412820	236180
SMALL	811991	1197994	1053744	1487452

Table 15(10)

Estimated Total Number of Students Who Are Eligible for Assessment by Tape Sample

Age 13

	TAPE 1	TAPE 2	TAPE 3	TAPE 4
TOTAL	3310355	3348263	3338947	3340216
SEX:				
MALE	1747611	1677070	1629646	1785519
FEMALE	1562742	1669346	1709301	1554697
RACE:				
WHITE	2470449	2501335	2476408	2478442
BLACK	459418	469832	480222	467216
HISPANIC	289113	282776	292047	299025
OTHER	91374	94320	90270	95533
REGION:				
NORTHEAST	761536	690015	731394	733441
SOUTHEAST	870751	955525	804766	870643
CENTRAL	856550	841877	794101	835296
WEST	821518	860846	1008687	900836
PARENTS ED:				
LESS THAN HIGH SCHOOL	307983	210519	330924	284487
HIGH SCHOOL	1226582	1161261	1283853	1257443
GREATER THAN HIGH SCHOOL	1494458	1575838	1431932	1489634
UNKNOWN	281330	400644	292237	308651
SIZE AND TYPE OF COMMUNITY:				
RURAL	257086	166458	145597	398820
DISADVANTAGED URBAN	352010	178032	361847	165746
ADVANTAGED URBAN	390384	201802	301008	316831
BIG CITY	294904	368039	211471	344823
FRINGE	588801	659332	721578	578886
MEDIUM	470814	648954	481058	670936
SMALL	956356	1125646	1116388	864175

Table 15(11)

Estimated Total Number of Students Who Are Eligible for Assessment by Tape Sample

Age 17

	TAPE 1	TAPE 2	TAPE 3	TAPE 4
TOTAL	3048025	3045283	2978520	3026687
SEX:				
MALE	1618433	1583931	1519193	1485595
FEMALE	1429591	1461350	1459326	1541091
RACE:				
WHITE	2286809	2277417	2252712	2261915
BLACK	419084	434916	425528	441416
HISPANIC	255867	248980	212199	247841
OTHER	86264	83970	88082	75514
REGION:				
NORTHEAST	739136	717514	647244	767227
SOUTHEAST	733342	823804	780987	704032
CENTRAL	784021	723271	758629	823867
WEST	791526	780695	791660	731561
PARENTS ED:				
LESS THAN HIGH SCHOOL	359225	361064	360424	313972
HIGH SCHOOL	1072167	1069930	1055548	1075779
GREATER THAN HIGH SCHOOL	1500148	1450475	1476708	1543271
UNKNOWN	116484	163813	85840	93664
SIZE AND TYPE OF COMMUNITY:				
RURAL	184826	75975	101252	217805
DISADVANTAGED URBAN	356393	299455	254331	301164
ADVANTAGED URBAN	534485	509189	296041	528053
BIG CITY	133848	225307	402884	215027
FRINGE	334566	223363	364522	287883
MEDIUM	582749	504779	510419	524017
SMALL	921158	1207215	1049070	952738

Table 15(12)a

Number of Students Receiving Reading and Writing Items and Plausible Values

Grade 4/Age 9

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	18945	20095	12953	26087
STUDENTS WITH READING:				
ITEMS	18497	19637	12660	25474
PLAUSIBLE VALUES	16799	17840	11507	23132
STUDENTS WITH WRITING:				
ITEMS	16025	16987	10986	22026
PLAUSIBLE VALUES	5795	8807	5795	8807

Table 15(12)b

Weighted Counts of Students Receiving Reading and Writing Items and Plausible Values

Grade 4/Age 9

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3067047	3200290	2295588	3971749
STUDENTS WITH READING:				
ITEMS	2991059	3125304	2240814	3875548
PLAUSIBLE VALUES	2710595	2837712	2031707	3516600
STUDENTS WITH WRITING:				
ITEMS	2600057	2709078	1951487	3357648
PLAUSIBLE VALUES	1026813	1408047	1026813	1408047

Table 15(13)a

Number of Students Receiving Reading and Writing Items and Plausible Values

Grade 8/Age 13

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	21070	21850	14515	28405
STUDENTS WITH READING:				
ITEMS	20568	21324	14173	27719
PLAUSIBLE VALUES	17535	18173	12043	23665
STUDENTS WITH WRITING:				
ITEMS	17810	18498	12289	24019
PLAUSIBLE VALUES	7420	11092	7420	11092

Table 15(13)b

Weighted Counts of Students Receiving Reading and Writing Items and Plausible Values

Grade 8/Age 13

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3319320	3313949	2269841	4363428
STUDENTS WITH READING:				
ITEMS	3241767	3235027	2217367	4259426
PLAUSIBLE VALUES	2763807	2761459	1888098	3637168
STUDENTS WITH WRITING:				
ITEMS	2805295	2802901	1918064	3690132
PLAUSIBLE VALUES	1155803	1682192	1155803	1682192

Table 15(14)a

Number of Students Receiving Reading and Writing Items and Plausible Values

Grade 11/Age 17

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	22783	22865	16787	28861
STUDENTS WITH READING:				
ITEMS	22226	22325	16381	28170
PLAUSIBLE VALUES	18984	19080	14009	24055
STUDENTS WITH WRITING:				
ITEMS	19267	19367	14219	24415
PLAUSIBLE VALUES	7919	10657	7919	10657

Table 15(14)b

Weighted Counts of Students Receiving Reading and Writing Items and Plausible Values

Grade 11/Age 17

	ELIGIBLE BY AGE	ELIGIBLE BY GRADE	ELIGIBLE BY AGE AND GRADE	ELIGIBLE BY AGE OR GRADE
TOTAL	3010157	3072622	2037738	4045040
STUDENTS WITH READING:				
ITEMS	2936218	2999803	1987958	3948063
PLAUSIBLE VALUES	2509020	2563822	1699683	3373158
STUDENTS WITH WRITING:				
ITEMS	2545582	2600027	1724920	3420689
PLAUSIBLE VALUES	963071	1430241	963071	1430241

NAEP 1983-84 READING AND WRITING ASSESSMENT - 4TH GRADERS Table 15(15)
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

GENERAL READING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	17840	2837712(0%)	217.4(0.7)	38.0(0.4)	168.4(0.9)	191.9(0.7)	218.0(0.8)	243.7(1.0)	266.0(1.1)
SEX									
MALE	9063	1412664(1%)	214.8(0.9)	39.3(0.5)	164.3(0.9)	188.1(0.9)	215.2(1.2)	242.4(1.3)	265.4(1.8)
FEMALE	8777	1425047(1%)	220.0(0.7)	36.6(0.5)	173.3(1.3)	195.8(0.8)	220.4(0.8)	244.8(1.1)	266.7(1.2)
ETHNICITY/RACE									
WHITE	11782	2002444(1%)	224.8(0.9)	36.1(0.4)	178.3(1.4)	200.8(0.9)	225.3(0.9)	249.4(1.1)	271.1(1.0)
BLACK	2796	431187(1%)	195.0(1.3)	35.3(0.8)	150.1(3.2)	171.3(1.5)	195.6(1.4)	218.4(1.4)	239.3(1.9)
HISPANIC	2469	322624(2%)	200.7(1.0)	36.9(0.7)	153.3(1.6)	176.2(1.6)	201.3(1.4)	226.3(1.3)	247.8(1.6)
OTHER	793	81456(4%)	221.1(1.8)	37.0(1.5)	172.6(2.8)	196.5(3.3)	222.0(1.4)	246.2(3.1)	268.6(2.0)
REGION									
NORTHEAST	4061	633342(2%)	220.8(1.6)	37.1(1.0)	172.9(2.7)	195.9(1.9)	221.2(2.0)	246.5(2.1)	268.7(2.3)
SOUTHEAST	4520	673333(5%)	212.6(1.6)	38.0(0.6)	163.9(1.0)	187.3(1.7)	212.9(2.0)	238.5(1.6)	260.9(2.0)
CENTRAL	4940	757011(5%)	221.2(1.8)	37.7(0.8)	172.4(3.3)	195.7(2.0)	222.1(1.6)	247.2(2.1)	269.8(2.3)
WEST	4319	774026(2%)	215.0(1.3)	38.5(0.9)	165.4(1.9)	189.3(2.0)	215.4(1.7)	241.7(1.5)	263.5(2.2)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1126	166134(5%)	200.0(1.2)	36.2(1.1)	153.8(2.6)	176.1(2.0)	200.7(1.1)	224.3(2.0)	245.1(1.3)
GRADUATED H.S.	3650	570097(3%)	215.4(0.8)	36.4(0.6)	168.3(1.7)	191.0(1.1)	216.5(1.1)	240.1(1.0)	260.9(0.8)
POST H.S.	6634	1075734(3%)	227.4(1.1)	38.2(0.5)	177.1(1.5)	202.0(1.0)	228.6(1.3)	254.0(1.6)	275.4(1.0)
UNKNOWN	6272	1001015(2%)	211.3(0.8)	36.3(0.5)	164.5(1.9)	187.1(0.8)	211.6(1.0)	236.1(0.9)	257.8(1.2)
SIZE/TYPE OF COMMUNITY									
RURAL	1158	183645(17%)	208.7(2.5)	38.6(0.8)	158.3(3.3)	182.0(3.0)	209.0(2.5)	235.6(3.1)	257.6(2.3)
DISADVANTAGED URBAN	2425	359699(16%)	197.9(1.5)	37.5(1.4)	149.5(2.7)	172.7(1.4)	197.5(1.3)	223.7(2.2)	246.6(3.7)
ADVANTAGED URBAN	2055	399737(15%)	234.7(2.3)	36.6(1.1)	188.2(2.5)	210.3(3.2)	234.8(2.6)	259.8(1.4)	281.2(2.1)
BIG CITIES	1345	214344(20%)	214.1(2.6)	36.7(0.9)	166.2(2.5)	189.1(3.8)	214.4(3.1)	239.9(2.8)	260.3(2.1)
FRINGE OF BIG CITIES	1841	313591(14%)	219.3(1.6)	35.6(0.7)	173.4(2.5)	195.3(1.7)	220.2(2.5)	243.6(2.2)	263.8(2.6)
MEDIUM CITIES	2755	457534(9%)	218.7(2.3)	36.2(0.7)	171.4(2.9)	194.5(2.4)	219.1(2.7)	243.7(2.6)	264.6(3.5)
SMALL PLACES	6261	909162(5%)	218.7(0.9)	36.7(0.5)	171.1(1.5)	193.9(1.3)	219.0(1.0)	243.8(1.1)	266.0(1.4)
AGE									
8 OR YOUNGER	130	17668(11%)	226.3(3.8)	38.8(2.7)	174.5(5.8)	199.9(8.1)	229.6(4.7)	254.7(15.1)	272.2(10.2)
9 YEARS OLD	11507	2031707(0%)	221.8(0.8)	36.8(0.4)	174.7(1.3)	197.3(1.2)	222.2(0.8)	247.0(1.2)	269.2(1.2)
10 OR OLDER	6203	788337(1%)	206.0(0.9)	30.9(15.5)	156.0(1.3)	179.7(1.2)	205.8(0.8)	232.4(1.1)	256.4(1.4)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

NAEP 1983-84 READING AND WRITING ASSESSMENT - 8TH GRADERS
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

Table 15(16)

GENERAL READING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	18173	2761459(0%)	260.6(0.5)	34.9(0.3)	215.6(0.7)	238.1(0.6)	261.3(0.6)	284.4(0.6)	304.9(0.8)
SEX									
MALE	9066	1380877(1%)	257.0(0.6)	35.2(0.4)	211.5(1.1)	234.0(0.8)	258.1(0.8)	281.4(0.9)	300.5(1.2)
FEMALE	9106	1380477(1%)	264.1(0.6)	34.3(0.4)	219.9(1.0)	241.4(0.8)	264.5(0.8)	287.1(0.6)	308.2(0.7)
ETHNICITY/RACE									
WHITE	12939	2044478(0%)	266.5(0.6)	33.2(0.4)	224.3(1.4)	244.6(0.8)	267.1(0.9)	288.8(0.8)	308.9(0.9)
BLACK	2555	398198(1%)	240.1(1.1)	33.0(0.6)	197.4(1.7)	218.4(1.0)	241.2(1.3)	262.2(1.5)	281.5(1.2)
HISPANIC	2043	241189(2%)	242.9(1.3)	33.8(0.8)	197.7(2.1)	220.3(2.5)	244.1(1.4)	265.6(1.4)	284.9(1.4)
OTHER	636	77593(3%)	264.3(1.6)	34.1(1.0)	220.1(3.1)	240.9(1.9)	264.6(3.2)	288.0(1.9)	307.4(4.6)
REGION									
NORTHEAST	4109	627702(2%)	262.4(1.0)	34.8(0.6)	217.7(1.0)	240.0(1.3)	263.5(1.5)	285.8(0.7)	306.1(0.9)
SOUTHEAST	4589	648087(6%)	259.8(1.4)	36.2(0.8)	213.4(2.0)	235.6(1.9)	260.5(1.7)	284.7(2.2)	306.1(2.1)
CENTRAL	5061	726560(5%)	262.1(1.2)	34.0(0.4)	217.9(2.0)	239.9(1.3)	262.5(1.4)	285.3(1.1)	305.4(2.2)
WEST	4414	759110(2%)	258.3(0.7)	34.7(0.6)	213.4(1.9)	235.7(1.1)	259.0(0.9)	281.8(1.0)	302.3(1.3)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1833	262530(5%)	244.4(0.7)	33.1(0.8)	201.9(0.9)	222.4(1.0)	245.2(1.4)	266.6(1.2)	286.1(1.5)
GRADUATED H.S.	6444	974445(3%)	255.6(0.7)	32.9(0.4)	213.5(1.3)	234.2(1.4)	256.2(0.7)	277.6(0.7)	297.2(0.7)
POST H.S.	8117	1261623(2%)	271.4(0.7)	33.2(0.4)	228.7(1.0)	249.9(1.1)	272.3(1.1)	294.0(0.8)	312.6(1.0)
UNKNOWN	1609	234214(5%)	241.5(1.1)	34.3(0.6)	197.3(1.5)	219.1(1.4)	242.0(1.3)	263.7(1.9)	284.7(1.3)
SIZE/TYPE OF COMMUNITY									
RURAL	1082	146728(21%)!	259.7(2.3)	33.3(0.8)	216.0(3.2)	237.9(3.0)	260.6(2.9)	282.6(3.7)	301.6(2.0)
DISADVANTAGED URBAN	1812	244034(17%)	241.9(1.9)	33.9(1.4)	198.7(3.6)	219.7(2.4)	242.8(1.9)	263.9(2.5)	284.6(2.6)
ADVANTAGED URBAN	1977	292740(22%)!	276.3(2.6)	32.2(0.7)	236.1(3.7)	254.9(2.5)	277.0(2.3)	297.9(3.2)	317.4(2.3)
BIG CITIES	1839	286912(32%)!	258.7(1.7)	33.2(1.1)	216.2(2.8)	237.7(2.4)	259.3(1.4)	280.3(3.3)	300.1(2.7)
FRINGE OF BIG CITIES	2475	456315(18%)	261.6(1.2)	34.5(0.7)	216.7(1.8)	239.0(1.5)	262.3(1.2)	285.1(0.8)	304.9(1.8)
MEDIUM CITIES	2504	420128(18%)	259.9(2.5)	35.5(0.9)	213.6(3.0)	236.6(3.7)	261.1(2.5)	284.5(1.9)	304.2(3.1)
SMALL PLACES	6484	914603(7%)	261.0(0.9)	34.3(0.4)	216.7(1.6)	238.7(1.1)	261.7(1.0)	284.2(0.9)	304.9(1.1)
AGE									
12 OR YOUNGER	154	21346(16%)	265.5(4.3)	35.4(2.1)	216.6(6.7)	243.1(5.3)	267.9(4.0)	289.3(7.9)	309.0(5.7)
13 YEARS OLD	12043	1888098(0%)	266.2(0.6)	33.2(0.4)	223.9(1.2)	244.5(0.9)	266.7(0.8)	288.4(0.8)	308.6(0.8)
14 OR OLDER	5976	852015(1%)	247.9(0.8)	28.3(14.2)	202.4(0.8)	224.5(1.0)	248.0(0.8)	272.0(1.0)	293.2(1.2)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

NAEP 1983-84 READING AND WRITING ASSESSMENT - 11TH GRADERS Table 15(17)
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

GENERAL READING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	19080	2563822(0%)	289.1(0.8)	38.9(0.3)	239.1(1.0)	264.1(0.9)	290.4(0.7)	315.9(0.9)	337.8(0.9)
SEX									
MALE	9443	1292364(2%)	284.1(1.0)	39.3(0.4)	232.3(1.7)	258.5(1.1)	285.5(0.9)	311.0(1.2)	333.9(1.1)
FEMALE	9637	1271457(2%)	294.2(0.9)	37.8(0.4)	245.1(1.6)	269.1(1.0)	295.1(1.0)	320.1(1.1)	342.2(1.4)
ETHNICITY/RACE									
WHITE	13914	1906547(0%)	295.8(0.9)	37.0(0.3)	248.7(1.2)	271.9(0.9)	296.7(1.0)	321.0(0.8)	342.4(1.2)
BLACK	2792	383493(1%)	266.6(1.8)	36.1(0.6)	219.8(3.1)	242.4(1.5)	266.7(2.6)	291.4(2.5)	312.6(2.5)
HISPANIC	1699	203453(2%)	269.2(2.0)	38.2(0.7)	213.9(1.6)	243.5(3.6)	270.3(2.6)	294.9(2.3)	317.9(2.4)
OTHER	675	70329(3%)	287.2(2.2)	40.9(1.4)	233.2(6.3)	261.1(4.0)	288.5(2.2)	315.3(2.3)	338.2(3.7)
REGION									
NORTHEAST	4318	628410(2%)	290.1(2.7)	39.6(0.6)	238.6(2.1)	264.2(3.1)	291.3(2.5)	317.7(2.9)	339.8(2.7)
SOUTHEAST	4873	564382(8%)	287.1(1.7)	39.9(0.6)	234.9(2.0)	260.4(2.0)	288.2(2.0)	314.7(1.7)	337.8(1.7)
CENTRAL	5303	702531(6%)	290.7(1.7)	37.7(0.8)	241.5(2.8)	266.5(1.8)	291.9(1.4)	316.5(1.6)	337.4(1.9)
WEST	4586	668498(2%)	288.3(0.8)	38.6(0.7)	238.6(1.3)	263.5(1.2)	289.2(1.3)	314.4(1.5)	336.8(1.1)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	2300	293458(5%)	269.5(1.2)	36.9(0.6)	221.4(2.2)	244.2(1.6)	269.9(1.4)	295.5(1.4)	317.1(1.1)
GRADUATED H.S.	6600	865215(3%)	281.3(0.7)	36.9(0.6)	233.2(1.7)	257.3(0.8)	282.4(0.9)	306.3(1.0)	327.3(1.0)
POST H.S.	9378	1301603(3%)	300.5(0.9)	36.7(0.4)	253.3(1.2)	276.8(0.8)	301.4(1.0)	325.3(0.8)	346.6(1.1)
UNKNOWN	596	78893(5%)	260.3(2.1)	37.3(1.5)	213.0(6.0)	236.1(4.1)	260.4(2.9)	285.5(3.5)	307.7(4.9)
SIZE/TYPE OF COMMUNITY									
RURAL	1217	137029(22%)!	284.6(3.2)	38.2(1.0)	235.6(3.2)	257.9(3.7)	285.4(3.3)	311.0(3.9)	333.0(2.4)
DISADVANTAGED URBAN	1958	272210(21%)!	266.7(2.5)	37.8(0.9)	217.8(3.7)	241.6(2.1)	266.9(2.6)	292.4(2.8)	314.4(3.5)
ADVANTAGED URBAN	2546	420188(16%)	300.6(3.0)	38.9(1.0)	250.4(6.0)	276.0(3.4)	302.1(3.2)	327.5(2.5)	348.8(2.4)
BIG CITIES	1782	224698(24%)!	290.1(2.4)	37.0(1.2)	242.4(6.1)	266.9(1.8)	291.3(2.0)	315.2(2.0)	336.0(2.9)
FRINGE OF BIG CITIES	1848	262515(26%)!	289.9(1.3)	36.6(0.6)	242.6(2.4)	265.6(1.9)	290.1(0.9)	315.0(1.4)	335.8(1.6)
MEDIUM CITIES	3210	429862(9%)	292.4(1.2)	38.2(0.8)	242.3(1.3)	267.3(1.6)	293.5(1.5)	318.4(1.6)	340.6(2.1)
SMALL PLACES	6519	817320(5%)	289.2(1.0)	37.9(0.4)	239.8(1.2)	264.9(1.0)	290.4(0.7)	314.9(1.1)	336.6(1.3)
AGE									
16 OR YOUNGER	1992	334011(6%)	299.8(1.4)	36.2(0.8)	253.3(2.5)	275.7(1.1)	300.6(2.3)	323.8(1.6)	346.2(2.5)
17 YEARS OLD	14009	1699683(0%)	294.8(0.7)	36.5(0.4)	248.0(1.4)	271.0(1.0)	295.6(0.6)	319.9(0.7)	340.9(0.8)
18 OR OLDER	3079	530128(3%)	264.1(1.3)	30.1(15.1)	215.9(1.7)	238.9(1.2)	264.5(1.0)	289.7(1.7)	311.5(1.3)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

NAEP 1983-84 READING AND WRITING ASSESSMENT - 4TH GRADERS Table 15(18)
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

A.R.M. WRITING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	8807	1408047(1%)	1.58(0.01)	0.41(0.00)	0.96(0.01)	1.29(0.01)	1.57(0.01)	1.90(0.01)	2.17(0.01)
SEX									
MALE	4410	694799(2%)	1.50(0.01)	0.40(0.01)	0.90(0.01)	1.22(0.02)	1.50(0.01)	1.79(0.02)	2.10(0.01)
FEMALE	4397	713248(1%)	1.66(0.01)	0.41(0.01)	1.04(0.02)	1.35(0.01)	1.64(0.01)	1.98(0.01)	2.21(0.01)
ETHNICITY/RACE									
WHITE	5931	1016633(1%)	1.63(0.01)	0.40(0.01)	1.03(0.02)	1.34(0.01)	1.62(0.01)	1.96(0.01)	2.19(0.01)
BLACK	1298	196446(2%)	1.38(0.02)	0.40(0.01)	0.82(0.02)	1.06(0.02)	1.39(0.01)	1.67(0.02)	1.99(0.03)
HISPANIC	1169	152335(5%)	1.46(0.02)	0.40(0.01)	0.87(0.02)	1.14(0.03)	1.45(0.02)	1.73(0.02)	2.07(0.03)
OTHER	409	42633(7%)	1.60(0.03)	0.40(0.02)	0.99(0.05)	1.32(0.04)	1.59(0.03)	1.92(0.04)	2.18(0.03)
REGION									
NORTHEAST	2021	316585(2%)	1.61(0.02)	0.41(0.01)	0.98(0.02)	1.31(0.02)	1.60(0.02)	1.94(0.02)	2.18(0.01)
SOUTHEAST	2225	331763(6%)	1.54(0.02)	0.41(0.01)	0.92(0.01)	1.26(0.01)	1.54(0.01)	1.85(0.03)	2.14(0.02)
CENTRAL	2457	384599(5%)	1.60(0.02)	0.40(0.01)	0.98(0.02)	1.31(0.02)	1.59(0.02)	1.92(0.02)	2.17(0.02)
WEST	2104	375101(3%)	1.57(0.01)	0.41(0.01)	0.94(0.02)	1.28(0.02)	1.56(0.01)	1.89(0.02)	2.16(0.02)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	526	77745(5%)	1.43(0.03)	0.39(0.02)	0.87(0.03)	1.14(0.05)	1.44(0.03)	1.70(0.03)	2.03(0.03)
GRADUATED H.S.	1793	275493(4%)	1.54(0.01)	0.40(0.01)	0.93(0.02)	1.26(0.02)	1.54(0.01)	1.85(0.02)	2.14(0.01)
POST H.S.	3364	549929(3%)	1.66(0.01)	0.41(0.01)	1.05(0.02)	1.36(0.01)	1.65(0.01)	2.00(0.01)	2.21(0.01)
UNKNOWN	3067	495435(2%)	1.53(0.01)	0.40(0.01)	0.92(0.01)	1.26(0.01)	1.52(0.01)	1.83(0.02)	2.13(0.01)
SIZE/TYPE OF COMMUNITY									
RURAL	559	86997(15%)	1.53(0.02)	0.40(0.02)	0.92(0.04)	1.24(0.05)	1.52(0.03)	1.82(0.05)	2.12(0.02)
DISADVANTAGED URBAN	1102	162691(17%)	1.42(0.02)	0.41(0.01)	0.84(0.02)	1.10(0.03)	1.43(0.02)	1.71(0.02)	2.05(0.02)
ADVANTAGED URBAN	1077	207850(15%)	1.70(0.02)	0.40(0.01)	1.13(0.06)	1.39(0.02)	1.69(0.03)	2.03(0.02)	2.24(0.02)
BIG CITIES	646	103743(22%)!	1.55(0.03)	0.41(0.02)	0.93(0.03)	1.28(0.02)	1.55(0.03)	1.87(0.04)	2.15(0.02)
FRINGE OF BIG CITIES	926	155625(14%)	1.60(0.02)	0.40(0.01)	0.98(0.03)	1.31(0.02)	1.59(0.03)	1.92(0.04)	2.17(0.02)
MEDIUM CITIES	1392	233230(8%)	1.59(0.02)	0.40(0.01)	0.96(0.02)	1.29(0.02)	1.58(0.02)	1.92(0.03)	2.17(0.02)
SMALL PLACES	3105	457911(5%)	1.58(0.01)	0.40(0.01)	0.97(0.01)	1.30(0.01)	1.57(0.01)	1.90(0.02)	2.16(0.01)
AGE									
8 OR YOUNGER	75	9566(15%)	1.55(0.07)	0.40(0.05)	0.96(0.11)	1.29(0.14)	1.55(0.07)	1.84(0.13)	2.17(0.10)
9 YEARS OLD	5795	1026813(1%)	1.60(0.01)	0.41(0.01)	0.98(0.01)	1.31(0.01)	1.59(0.01)	1.93(0.01)	2.18(0.01)
10 OR OLDER	2937	371667(2%)	1.52(0.01)	0.37(0.12)	0.90(0.01)	1.23(0.02)	1.52(0.01)	1.82(0.03)	2.13(0.01)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

NAEP 1983-84 READING AND WRITING ASSESSMENT - 8TH GRADERS Table 15(19)
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

A.R.M. WRITING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	11092	1682192(1%)	2.05(0.01)	0.40(0.00)	1.45(0.01)	1.78(0.01)	2.04(0.01)	2.35(0.01)	2.64(0.01)
SEX									
MALE	5486	839776(1%)	1.96(0.01)	0.39(0.00)	1.38(0.01)	1.68(0.01)	1.97(0.01)	2.23(0.01)	2.56(0.01)
FEMALE	5606	842416(1%)	2.14(0.01)	0.39(0.00)	1.56(0.01)	1.85(0.01)	2.12(0.01)	2.45(0.01)	2.69(0.01)
ETHNICITY/RACE									
WHITE	7916	1249837(1%)	2.11(0.01)	0.39(0.00)	1.52(0.01)	1.83(0.01)	2.09(0.01)	2.41(0.01)	2.67(0.01)
BLACK	1500	232931(2%)	1.86(0.01)	0.38(0.01)	1.32(0.01)	1.56(0.02)	1.88(0.02)	2.14(0.02)	2.42(0.04)
HISPANIC	1271	150212(4%)	1.87(0.02)	0.39(0.01)	1.32(0.02)	1.55(0.02)	1.89(0.02)	2.16(0.02)	2.47(0.03)
OTHER	405	49212(4%)	2.09(0.03)	0.39(0.02)	1.49(0.04)	1.82(0.03)	2.09(0.04)	2.41(0.05)	2.67(0.03)
REGION									
NORTHEAST	2518	382006(2%)	2.09(0.01)	0.40(0.01)	1.49(0.01)	1.81(0.01)	2.08(0.01)	2.40(0.02)	2.66(0.01)
SOUTHEAST	2759	388695(7%)	2.03(0.02)	0.41(0.01)	1.42(0.02)	1.76(0.02)	2.02(0.01)	2.33(0.03)	2.64(0.02)
CENTRAL	3089	443166(5%)	2.06(0.01)	0.39(0.01)	1.46(0.01)	1.79(0.01)	2.05(0.01)	2.35(0.02)	2.63(0.01)
WEST	2726	468325(2%)	2.03(0.02)	0.40(0.01)	1.43(0.01)	1.76(0.01)	2.03(0.01)	2.33(0.03)	2.62(0.02)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1072	152336(5%)	1.89(0.02)	0.38(0.01)	1.34(0.02)	1.59(0.03)	1.91(0.02)	2.17(0.02)	2.48(0.03)
GRADUATED H.S.	3887	584033(3%)	2.02(0.01)	0.38(0.01)	1.43(0.01)	1.77(0.01)	2.02(0.01)	2.29(0.01)	2.60(0.01)
POST H.S.	5038	785037(3%)	2.13(0.01)	0.39(0.00)	1.54(0.01)	1.84(0.01)	2.12(0.01)	2.45(0.01)	2.69(0.01)
UNKNOWN	1000	144685(5%)	1.90(0.02)	0.39(0.02)	1.34(0.01)	1.59(0.02)	1.91(0.02)	2.18(0.03)	2.49(0.04)
SIZE/TYPE OF COMMUNITY									
RURAL	645	86112(21%)!	2.03(0.03)	0.38(0.02)	1.44(0.04)	1.77(0.03)	2.02(0.02)	2.30(0.03)	2.61(0.02)
DISADVANTAGED URBAN	1073	142985(18%)	1.88(0.02)	0.38(0.01)	1.33(0.02)	1.57(0.03)	1.90(0.02)	2.16(0.02)	2.45(0.04)
ADVANTAGED URBAN	1229	181317(22%)!	2.21(0.02)	0.39(0.01)	1.68(0.05)	1.90(0.02)	2.19(0.02)	2.52(0.02)	2.73(0.02)
BIG CITIES	1130	177473(34%)!	2.01(0.02)	0.39(0.01)	1.42(0.02)	1.75(0.04)	2.02(0.02)	2.30(0.05)	2.60(0.02)
FRINGE OF BIG CITIES	1536	282827(18%)	2.07(0.02)	0.39(0.00)	1.47(0.02)	1.80(0.02)	2.06(0.02)	2.37(0.03)	2.64(0.01)
MEDIUM CITIES	1551	257076(18%)	2.04(0.03)	0.40(0.01)	1.43(0.02)	1.77(0.02)	2.03(0.02)	2.34(0.04)	2.64(0.02)
SMALL PLACES	3928	554402(7%)	2.05(0.01)	0.39(0.01)	1.45(0.01)	1.79(0.01)	2.05(0.01)	2.35(0.02)	2.64(0.01)
AGE									
12 OR YOUNGER	89	12763(19%)	2.07(0.06)	0.37(0.03)	1.54(0.13)	1.84(0.04)	2.07(0.06)	2.35(0.11)	2.63(0.05)
13 YEARS OLD	7420	1155803(1%)	2.08(0.01)	0.39(0.00)	1.48(0.01)	1.81(0.01)	2.07(0.01)	2.39(0.01)	2.66(0.01)
14 OR OLDER	3583	513626(2%)	1.98(0.01)	0.33(0.17)	1.38(0.01)	1.68(0.02)	1.98(0.01)	2.25(0.01)	2.58(0.01)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

NAEP 1983-84 READING AND WRITING ASSESSMENT - 11TH GRADERS Table 15(20)
WEIGHTED MEANS, STANDARD DEVIATION(N-1), AND PERCENTILES FOR REPORTING GROUPS

A.R.M. WRITING PROFICIENCY (AVERAGE OF 5 PLAUSIBLE VALUES)

	N	WEIGHTED N	MEAN	ST. DEV.	- 10 -	- 25 -	- 50 -	- 75 -	- 90 -
-- TOTAL --	10657	1430241(1%)	2.19(0.01)	0.44(0.01)	1.55(0.02)	1.87(0.01)	2.18(0.01)	2.53(0.01)	2.75(0.02)
SEX									
MALE	5215	714418(2%)	2.09(0.01)	0.43(0.01)	1.45(0.01)	1.79(0.01)	2.08(0.01)	2.42(0.01)	2.68(0.01)
FEMALE	5442	715823(2%)	2.29(0.01)	0.43(0.01)	1.73(0.04)	1.95(0.01)	2.30(0.02)	2.61(0.01)	2.91(0.02)
ETHNICITY/RACE									
WHITE	7892	1077899(1%)	2.24(0.01)	0.43(0.01)	1.65(0.03)	1.92(0.01)	2.24(0.01)	2.57(0.01)	2.83(0.02)
BLACK	1478	205670(2%)	2.00(0.02)	0.41(0.01)	1.39(0.02)	1.70(0.04)	2.00(0.02)	2.30(0.04)	2.62(0.02)
HISPANIC	902	107250(3%)	2.00(0.02)	0.42(0.01)	1.37(0.02)	1.68(0.04)	2.00(0.02)	2.31(0.04)	2.62(0.03)
OTHER	385	39422(4%)	2.16(0.03)	0.44(0.03)	1.52(0.05)	1.84(0.03)	2.14(0.04)	2.49(0.03)	2.73(0.04)
REGION									
NORTHEAST	2459	354526(2%)	2.22(0.03)	0.44(0.01)	1.58(0.05)	1.89(0.02)	2.22(0.03)	2.56(0.02)	2.81(0.05)
SOUTHEAST	2705	313707(9%)	2.16(0.02)	0.44(0.01)	1.50(0.03)	1.84(0.02)	2.15(0.02)	2.51(0.02)	2.73(0.01)
CENTRAL	2959	390762(7%)	2.20(0.02)	0.43(0.01)	1.53(0.04)	1.88(0.02)	2.19(0.02)	2.54(0.02)	2.78(0.05)
WEST	2534	371246(3%)	2.17(0.01)	0.44(0.01)	1.53(0.02)	1.85(0.01)	2.16(0.01)	2.51(0.01)	2.74(0.01)
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1267	159736(5%)	1.99(0.02)	0.43(0.01)	1.36(0.02)	1.67(0.04)	1.99(0.02)	2.28(0.04)	2.61(0.02)
GRADUATED H.S.	3675	479173(3%)	2.15(0.01)	0.42(0.01)	1.52(0.02)	1.84(0.01)	2.14(0.01)	2.49(0.01)	2.72(0.01)
POST H.S.	5312	740485(3%)	2.27(0.01)	0.43(0.01)	1.69(0.04)	1.93(0.01)	2.27(0.01)	2.60(0.01)	2.88(0.02)
UNKNOWN	290	37087(7%)	1.99(0.03)	0.41(0.02)	1.38(0.04)	1.70(0.07)	2.00(0.04)	2.28(0.07)	2.60(0.03)
SIZE/TYPE OF COMMUNITY									
RURAL	699	79285(22%)!	2.13(0.03)	0.44(0.01)	1.48(0.04)	1.82(0.04)	2.12(0.05)	2.48(0.04)	2.72(0.02)
DISADVANTAGED URBAN	1029	142348(22%)!	2.01(0.02)	0.41(0.02)	1.40(0.02)	1.72(0.05)	2.01(0.02)	2.30(0.05)	2.62(0.02)
ADVANTAGED URBAN	1458	240121(16%)	2.28(0.02)	0.44(0.01)	1.70(0.05)	1.95(0.02)	2.30(0.02)	2.62(0.02)	2.92(0.04)
BIG CITIES	996	126101(24%)!	2.18(0.02)	0.43(0.01)	1.55(0.04)	1.86(0.02)	2.16(0.03)	2.51(0.03)	2.75(0.03)
FRINGE OF BIG CITIES	1011	142241(27%)!	2.19(0.02)	0.43(0.01)	1.57(0.03)	1.88(0.02)	2.19(0.03)	2.53(0.02)	2.73(0.01)
MEDIUM CITIES	1807	241537(9%)	2.21(0.02)	0.43(0.02)	1.58(0.04)	1.88(0.02)	2.20(0.03)	2.55(0.02)	2.78(0.05)
SMALL PLACES	3657	458608(5%)	2.19(0.01)	0.43(0.01)	1.56(0.02)	1.87(0.01)	2.19(0.02)	2.53(0.02)	2.76(0.04)
AGE									
16 OR YOUNGER	1102	184232(6%)	2.23(0.02)	0.44(0.01)	1.62(0.05)	1.90(0.02)	2.23(0.03)	2.57(0.02)	2.82(0.06)
17 YEARS OLD	7919	963071(1%)	2.21(0.01)	0.43(0.01)	1.58(0.02)	1.89(0.01)	2.20(0.01)	2.55(0.01)	2.79(0.03)
18 OR OLDER	1636	282938(3%)	2.08(0.02)	0.36(0.18)	1.43(0.02)	1.78(0.02)	2.08(0.02)	2.43(0.02)	2.68(0.02)

! INTERPRET WITH CAUTION. STANDARD ERRORS ARE POORLY ESTIMATED.

Table 15(21)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 4	%-OMIT
-- TOTAL --	17840	2837712(0%)	100.0(0.0) 217.5(0.7)	0.0
SEX				
MALE	9063	1412664(1%)	100.0(0.0) 215.1(0.9)	0.0
FEMALE	8777	1425047(1%)	100.0(0.0) 220.0(0.7)	0.0
ETHNICITY/RACE				
WHITE	11782	2002444(1%)	100.0(0.0) 224.9(0.9)	0.0
BLACK	2796	431187(1%)	100.0(0.0) 194.9(1.3)	0.0
HISPANIC	2469	322624(2%)	100.0(0.0) 201.2(1.0)	0.0
OTHER	793	81456(4%)	100.0(0.0) 221.6(1.8)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	1126	166134(5%)	100.0(0.0) 200.2(1.2)	0.0
GRADUATED H.S.	3650	570097(3%)	100.0(0.0) 215.5(0.8)	0.0
POST H.S.	6634	1075734(3%)	100.0(0.0) 227.4(1.1)	0.0
UNKNOWN	6272	1001015(2%)	100.0(0.0) 211.6(0.8)	0.0
AGE				
9 YEARS OLD	11507	2031707(0%)	100.0(0.0) 221.7(0.8)	0.0
10 OR OLDER	6203	788337(1%)	100.0(0.0) 206.5(0.9)	0.0

Table 15(22)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	17840	2837712(0%)	49.8(0.5) 215.1(0.9)	50.2(0.5) 220.0(0.7)	0.0
SEX					
MALE	9063	1412664(1%)	100.0(0.0) 215.1(0.9)	0.0(0.0) *****(0.0)	0.0
FEMALE	8777	1425047(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 220.0(0.7)	0.0
ETHNICITY/RACE					
WHITE	11782	2002444(1%)	49.4(0.6) 222.8(1.1)	50.6(0.6) 226.9(0.9)	0.0
BLACK	2796	431187(1%)	47.6(1.2) 191.3(1.7)	52.4(1.2) 198.2(1.5)	0.0
HISPANIC	2469	322624(2%)	53.8(1.3) 198.9(1.3)	46.2(1.3) 204.0(1.4)	0.0
OTHER	793	81456(4%)	54.8(1.8) 215.7(2.1)	45.2(1.8) 228.7(2.5)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	1126	166134(5%)	45.5(1.9) 195.1(1.9)	54.5(1.9) 204.5(1.7)	0.0
GRADUATED H.S.	3650	570097(3%)	50.0(0.9) 211.5(1.1)	50.0(0.9) 219.6(1.2)	0.0
POST H.S.	6634	1075734(3%)	52.1(0.7) 224.8(1.4)	47.9(0.7) 230.2(1.0)	0.0
UNKNOWN	6272	1001015(2%)	47.7(0.9) 209.7(0.9)	52.3(0.9) 213.3(1.1)	0.0
AGE					
9 YEARS OLD	11507	2031707(0%)	46.5(0.6) 220.2(1.1)	53.5(0.6) 223.1(0.8)	0.0
10 OR OLDER	6203	788337(1%)	58.3(0.8) 204.3(1.1)	41.7(0.8) 209.5(1.3)	0.0

Table 15(23)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	17840	2837712(0%)	70.6(0.2) 224.9(0.9)	15.2(0.1) 194.9(1.3)	11.4(0.2) 201.2(1.0)	1.3(0.1) 216.4(2.5)	1.5(0.1) 226.0(2.9)	0.0(0.0) 219.7(12.7)	0.0
SEX									
MALE	9063	1412664(1%)	70.0(0.5) 222.8(1.1)	14.5(0.4) 191.3(1.7)	12.3(0.4) 198.9(1.3)	1.6(0.1) 211.2(3.0)	1.6(0.2) 219.9(3.3)	0.0(0.0) 228.9(28.4)	0.0
FEMALE	8777	1425047(1%)	71.1(0.5) 226.9(0.9)	15.9(0.4) 198.2(1.5)	10.5(0.3) 204.0(1.4)	1.0(0.1) 224.2(3.3)	1.5(0.1) 232.2(3.6)	0.0(0.0) 211.3(11.7)	0.0
ETHNICITY/RACE									
WHITE	11782	2002444(1%)	100.0(0.0) 224.9(0.9)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	2796	431187(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 194.9(1.3)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	2469	322624(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 201.2(1.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	793	81456(4%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	45.3(3.4) 216.4(2.5)	53.5(3.3) 226.0(2.9)	1.1(0.5) 219.7(12.7)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1126	166134(5%)	61.2(2.2) 207.9(1.4)	17.4(1.4) 184.2(4.2)	18.8(1.7) 189.2(3.5)	1.9(0.4) 205.1(9.0)	0.8(0.2) 207.1(18.3)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3650	570097(3%)	71.6(1.0) 222.3(0.9)	16.0(0.8) 195.2(2.0)	10.5(0.7) 200.3(1.8)	1.3(0.2) 217.9(4.6)	0.5(0.1) 220.4(6.3)	0.0(0.0) *****(0.0)	0.0
POST H.S.	6634	1075734(3%)	72.7(0.6) 235.1(1.2)	14.4(0.5) 199.8(1.8)	9.8(0.4) 209.7(1.7)	1.3(0.1) 221.3(3.8)	1.8(0.2) 236.4(3.1)	0.0(0.0) 219.4(30.1)	0.0
UNKNOWN	6272	1001015(2%)	69.9(0.7) 217.8(1.0)	15.0(0.6) 192.3(2.1)	11.9(0.6) 198.4(1.4)	1.2(0.1) 213.3(4.7)	1.9(0.3) 218.1(3.5)	0.1(0.0) 219.9(18.7)	0.0
AGE									
9 YEARS OLD	11507	2031707(0%)	72.4(0.3) 228.1(1.0)	13.9(0.1) 199.1(1.5)	10.8(0.3) 207.0(1.3)	1.2(0.1) 222.8(2.8)	1.6(0.1) 229.0(3.2)	0.0(0.0) 222.0(13.5)	0.0
10 OR OLDER	6203	788337(1%)	66.0(0.3) 215.4(1.0)	18.3(0.3) 186.6(1.8)	12.9(0.4) 188.7(1.4)	1.6(0.1) 203.5(3.5)	1.2(0.2) 213.7(3.9)	0.0(0.0) 199.3(****)	0.0

Table 15(24)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	17840	2837712(0%)	22.3(0.4) 221.6(1.6)	23.7(1.3) 212.2(1.6)	26.7(1.3) 221.5(1.8)	27.3(0.5) 215.0(1.3)	0.0
SEX							
MALE	9063	1412664(1%)	22.3(0.6) 220.5(2.0)	23.4(1.4) 208.3(2.0)	26.4(1.5) 218.6(2.3)	27.9(0.7) 213.1(1.4)	0.0
FEMALE	8777	1425047(1%)	22.4(0.7) 222.7(1.9)	24.0(1.3) 216.0(1.4)	26.9(1.3) 224.2(1.3)	26.7(0.8) 217.0(1.7)	0.0
ETHNICITY/RACE							
WHITE	11782	2002444(1%)	23.5(0.2) 227.5(1.7)	21.1(1.5) 222.0(1.7)	30.7(1.5) 225.7(1.6)	24.7(0.2) 223.8(2.3)	0.0
BLACK	2796	431187(1%)	21.2(0.4) 199.6(3.6)	44.6(0.5) 192.2(1.7)	19.4(3.4) 196.5(3.0)	14.8(3.5) 194.2(3.3)	0.0
HISPANIC	2469	322624(2%)	17.6(3.3) 208.7(3.1)	14.7(4.4) 204.1(2.6)	12.9(2.6) 207.5(2.9)	54.8(1.0) 196.6(1.1)	0.0
OTHER	793	81456(4%)	17.1(2.7) 221.8(2.5)	14.7(3.0) 219.4(6.3)	22.0(3.0) 222.9(3.8)	46.2(5.5) 221.6(3.1)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1126	166134(5%)	17.1(1.5) 204.9(4.5)	32.9(2.6) 197.3(2.4)	22.3(2.4) 203.5(1.6)	27.6(3.0) 198.2(2.6)	0.0
GRADUATED H.S.	3650	570097(3%)	22.2(1.6) 221.2(1.8)	25.9(1.9) 209.0(1.9)	29.6(2.1) 219.6(1.4)	22.4(1.1) 212.0(2.1)	0.0
POST H.S.	6634	1075734(3%)	22.1(1.1) 231.1(2.0)	22.6(1.5) 222.9(2.6)	26.6(1.9) 230.0(2.2)	28.7(1.3) 225.7(1.7)	0.0
UNKNOWN	6272	1001015(2%)	23.6(1.0) 214.9(2.1)	22.3(1.7) 206.7(1.2)	26.1(1.6) 216.2(1.6)	28.0(1.1) 208.5(1.4)	0.0
AGE							
9 YEARS OLD	11507	2031707(0%)	22.5(0.4) 225.4(1.5)	23.4(1.4) 217.8(1.9)	27.0(1.4) 225.0(1.9)	27.0(0.5) 218.9(1.6)	0.0
10 OR OLDER	6203	708337(1%)	21.3(0.5) 211.5(3.0)	24.7(1.4) 198.2(1.3)	25.9(1.5) 211.5(2.0)	28.1(0.6) 205.2(1.1)	0.0

Table 15(25)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT AGE

	N	WEIGHTED N	7-LESS	8	9	10	11	12-MORE	%-OMIT
-- TOTAL --	17840	2837712(0%)	0.0(0.0) 182.7(27.0)	0.6(0.1) 229.2(3.9)	71.6(0.1) 221.7(0.8)	24.5(0.2) 209.0(0.9)	2.9(0.2) 188.3(1.6)	0.4(0.1) 181.5(6.5)	0.0
SEX									
MALE	9063	1412664(1%)	0.0(0.0) 165.5(16.2)	0.5(0.1) 234.7(6.1)	66.9(0.4) 220.2(1.1)	28.3(0.4) 207.0(1.2)	3.7(0.3) 187.2(2.4)	0.5(0.1) 179.1(5.7)	0.0
FEMALE	8777	1425047(1%)	0.0(0.0) 223.7(****)	0.7(0.1) 225.0(4.7)	76.2(0.4) 223.1(0.8)	20.8(0.3) 211.7(1.4)	2.1(0.2) 190.3(2.4)	0.2(0.1) 187.9(10.2)	0.0
ETHNICITY/RACE									
WHITE	11782	2002444(1%)	0.0(0.0) *****(0.0)	0.5(0.1) 241.0(4.7)	73.5(0.1) 228.1(1.0)	23.9(0.2) 217.1(1.1)	1.9(0.2) 196.5(1.9)	0.2(0.0) 189.7(7.2)	0.0
BLACK	2796	431187(1%)	0.0(0.0) 158.4(14.6)	0.8(0.2) 200.6(4.8)	65.6(0.4) 199.1(1.5)	26.9(0.7) 188.1(1.8)	6.1(0.6) 181.1(4.1)	0.5(0.2) 177.5(7.8)	0.0
HISPANIC	2469	322624(2%)	0.0(0.0) 174.5(****)	0.6(0.1) 208.1(9.7)	68.0(0.8) 207.0(1.3)	25.8(1.0) 191.0(1.2)	4.8(0.5) 179.5(4.4)	0.9(0.3) 172.2(22.1)	0.0
OTHER	793	81456(4%)	0.1(0.1) 223.7(****)	1.5(0.5) 238.4(13.7)	71.8(1.3) 226.3(2.2)	22.3(1.1) 211.0(2.9)	3.8(0.6) 193.2(6.5)	0.5(0.2) 172.5(10.6)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1126	166134(5%)	0.0(0.0) *****(0.0)	0.1(0.1) 203.4(21.1)	60.9(1.9) 205.1(1.1)	31.5(1.9) 195.6(2.3)	7.0(0.7) 181.2(4.0)	0.4(0.2) 161.1(44.6)	0.0
GRADUATED H.S.	3650	570097(3%)	0.0(0.0) *****(0.0)	0.4(0.1) 209.5(13.3)	69.3(0.9) 219.4(0.9)	27.0(0.8) 209.2(1.5)	3.0(0.3) 187.7(3.3)	0.3(0.1) 189.3(6.0)	0.0
POST H.S.	6634	1075734(3%)	0.0(0.0) 165.5(16.2)	0.8(0.1) 239.0(5.4)	74.9(0.5) 231.2(1.2)	22.1(0.4) 217.6(1.5)	1.9(0.2) 193.3(3.4)	0.2(0.1) 179.1(10.9)	0.0
UNKNOWN	6272	1001015(2%)	0.0(0.0) 223.7(****)	0.6(0.1) 222.0(5.9)	71.4(0.6) 215.3(0.9)	24.4(0.5) 204.1(1.2)	3.2(0.3) 188.2(3.0)	0.4(0.1) 182.0(5.9)	0.0
AGE									
9 YEARS OLD	11507	2031707(0%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 221.7(0.8)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
10 OR OLDER	6203	788337(1%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	88.2(0.7) 209.0(0.9)	10.5(0.6) 188.3(1.6)	1.3(0.2) 181.5(6.5)	0.0

Table 15(26)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	17840	2837712(0%)	6.5(1.1) 207.7(2.5)	12.7(2.0) 198.5(1.5)	14.1(2.1) 234.5(2.3)	7.6(1.5) 214.2(2.6)	11.1(1.6) 219.2(1.6)	16.1(1.4) 218.9(2.3)	32.0(1.7) 219.2(0.9)	0.0
SEX										
MALE	9063	1412664(1%)	6.6(1.1) 204.6(2.5)	12.2(2.0) 195.1(1.9)	14.3(2.1) 233.0(3.5)	7.8(1.5) 211.0(2.9)	11.4(1.8) 216.6(1.8)	16.3(1.5) 217.2(3.1)	31.5(1.8) 216.3(1.2)	0.0
FEMALE	8777	1425047(1%)	6.3(1.2) 210.9(3.4)	13.2(2.2) 201.6(1.6)	13.9(2.1) 236.0(2.2)	7.3(1.6) 217.6(2.8)	10.7(1.4) 221.9(1.8)	16.0(1.3) 220.6(1.7)	32.6(1.7) 221.9(0.9)	0.0
ETHNICITY/RACE										
WHITE	11782	2002444(1%)	6.3(1.1) 217.1(2.2)	6.5(1.7) 212.8(2.7)	16.1(2.5) 237.2(2.1)	6.2(1.6) 222.5(2.8)	12.2(1.6) 222.6(1.4)	16.4(1.3) 225.7(2.2)	36.3(1.9) 223.7(0.9)	0.0
BLACK	2796	431187(1%)	5.2(1.6) 181.1(3.3)	35.9(6.0) 190.6(1.7)	7.1(2.0) 217.8(6.3)	10.2(3.5) 197.3(3.3)	4.6(1.3) 198.4(5.9)	13.6(2.6) 194.6(2.7)	23.3(3.6) 196.1(1.9)	0.0
HISPANIC	2469	322624(2%)	10.0(5.1) 190.1(3.4)	20.5(5.4) 187.5(2.5)	9.5(2.4) 221.5(4.1)	11.4(3.1) 203.8(2.3)	11.0(3.6) 205.4(3.4)	17.8(4.6) 204.4(1.5)	19.8(2.9) 204.7(2.4)	0.0
OTHER	793	81456(4%)	2.8(1.0) 197.4(5.6)	11.5(1.8) 207.0(4.5)	20.7(2.8) 236.1(4.7)	12.0(2.5) 224.3(4.7)	17.2(4.3) 224.7(3.2)	14.6(2.2) 220.2(5.2)	21.3(2.9) 215.4(3.7)	0.0
PARENTAL EDUCATION										
NOT GRADUATED H.S.	1126	166134(5%)	10.3(2.4) 189.4(3.1)	14.4(2.6) 188.4(3.8)	4.6(1.1) 214.6(6.1)	6.8(1.7) 198.4(4.0)	7.3(1.1) 214.1(4.8)	21.0(2.9) 204.8(3.2)	35.7(3.0) 201.1(1.8)	0.0
GRADUATED H.S.	3650	570097(3%)	9.3(1.6) 212.0(3.1)	11.1(2.1) 199.0(3.2)	6.4(1.0) 225.0(3.9)	6.5(1.6) 213.7(2.9)	9.5(1.3) 219.1(2.1)	15.2(1.5) 216.4(1.9)	42.0(2.7) 218.5(1.3)	0.0
POST H.S.	6634	1075734(3%)	4.8(0.7) 214.9(3.9)	11.2(1.8) 204.6(2.4)	21.3(3.2) 240.5(2.5)	7.5(1.5) 223.1(3.4)	11.4(2.0) 226.5(2.0)	15.3(1.3) 229.3(3.7)	28.5(1.6) 229.2(1.0)	0.0
UNKNOWN	6272	1001015(2%)	5.8(1.2) 204.3(3.6)	14.7(2.5) 195.5(1.8)	12.5(2.1) 227.7(2.7)	8.5(1.8) 208.3(2.9)	12.2(1.7) 213.0(2.4)	16.9(1.7) 213.1(1.9)	29.4(1.8) 213.8(1.3)	0.0
AGE										
9 YEARS OLD	11507	2031707(0%)	6.2(1.0) 212.2(2.3)	12.5(2.0) 203.6(1.8)	15.1(2.4) 237.6(2.2)	7.9(1.8) 218.4(2.8)	11.1(1.8) 221.7(1.9)	15.9(1.2) 223.0(3.0)	31.4(1.6) 223.4(0.9)	0.0
10 OR OLDER	6203	788337(1%)	7.4(1.6) 197.9(3.7)	13.0(2.3) 185.9(2.7)	11.4(1.5) 223.0(3.3)	6.4(1.1) 200.7(3.1)	10.8(1.2) 212.1(2.2)	16.7(2.3) 208.2(1.5)	34.3(2.4) 209.1(1.4)	0.0

Table 15(27)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	17682	2812980(1%)	5.9(0.3) 200.2(1.2)	20.3(0.6) 215.5(0.8)	38.2(0.9) 227.4(1.1)	35.6(0.7) 211.6(0.8)	0.9
SEX							
MALE	8965	1398027(1%)	5.4(0.3) 195.1(1.9)	20.4(0.7) 211.5(1.1)	40.1(1.1) 224.8(1.4)	34.1(0.9) 209.7(0.9)	1.0
FEMALE	8717	1414953(1%)	6.4(0.4) 204.5(1.7)	20.1(0.7) 219.6(1.2)	36.4(1.0) 230.2(1.0)	37.0(0.8) 213.3(1.1)	0.7
ETHNICITY/RACE							
WHITE	11717	1991399(1%)	5.1(0.3) 207.9(1.4)	20.5(0.7) 222.3(0.9)	39.3(1.1) 235.1(1.2)	35.1(0.8) 217.8(1.0)	0.6
BLACK	2757	425333(1%)	6.8(0.5) 184.2(4.2)	21.5(1.3) 195.2(2.0)	36.4(1.1) 199.8(1.8)	35.3(1.5) 192.3(2.1)	1.4
HISPANIC	2426	315870(3%)	9.9(1.1) 189.2(3.5)	19.0(1.2) 200.3(1.8)	33.4(1.5) 209.7(1.7)	37.7(1.6) 198.4(1.4)	2.1
OTHER	782	80377(4%)	5.5(1.0) 205.6(8.2)	13.0(1.2) 218.6(4.2)	41.3(2.9) 230.1(2.5)	40.2(2.8) 216.3(2.9)	1.3
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1126	166134(5%)	100.0(0.0) 200.2(1.2)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3650	570897(3%)	0.0(0.0) *****(0.0)	100.0(0.0) 215.5(0.8)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	6634	1075734(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 227.4(1.1)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	6272	1001015(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 211.6(0.8)	0.0
AGE							
9 YEARS OLD	11420	2016507(1%)	5.0(0.2) 205.1(1.1)	19.6(0.8) 219.4(0.9)	40.0(1.1) 231.2(1.2)	35.4(0.8) 215.3(0.9)	0.7
10 OR OLDER	6133	778997(1%)	8.3(0.6) 192.6(2.0)	22.2(0.6) 206.8(1.4)	33.5(0.9) 215.2(1.4)	36.0(0.9) 201.9(1.1)	1.2

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	150	200	250	300	350
-- TOTAL --	17840	2837712(0%)	96.2(0.2)	68.3(0.7)	19.8(0.7)	1.2(0.1)	0.0(0.0)
SEX							
MALE	9063	1412664(1%)	95.3(0.3)	65.1(0.9)	18.9(0.8)	1.1(0.1)	0.0(0.0)
FEMALE	8777	1425047(1%)	97.1(0.3)	71.5(0.7)	20.7(0.7)	1.2(0.2)	0.0(0.0)
ETHNICITY/RACE							
WHITE	11782	2002444(1%)	98.1(0.2)	75.8(0.8)	24.6(0.9)	1.5(0.2)	0.0(0.0)
BLACK	2796	431187(1%)	90.3(1.1)	44.7(1.5)	6.0(0.6)	0.0(0.0)	0.0(0.0)
HISPANIC	2469	322624(2%)	92.1(0.7)	52.2(1.4)	8.2(0.8)	0.1(0.1)	0.0(0.0)
OTHER	793	81456(4%)	97.4(0.5)	73.9(1.9)	20.3(2.4)	2.0(0.7)	0.0(0.0)
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1126	166134(5%)	92.5(0.9)	52.0(1.4)	7.0(0.8)	0.3(0.2)	0.0(0.0)
GRADUATED H.S.	3650	570097([illegible]%)	96.1(0.3)	68.1(1.0)	16.3(0.6)	0.7(0.2)	0.0(0.0)
POST H.S.	6634	1075734(3%)	97.6(0.3)	76.9(0.9)	28.6(1.2)	2.3(0.2)	0.0(0.0)
UNKNOWN	[illegible]	1001015(2%)	95.5(0.4)	62.4(1.0)	14.7(0.6)	0.3(0.1)	0.0(0.0)
AGE							
9 YEARS OLD	11507	2031707(0%)	97.4(0.2)	72.9(0.8)	22.4(0.9)	1.2(0.1)	0.0(0.0)
10 OR OLDER	6203	788337(1%)	93.1(0.5)	56.2(0.8)	12.6(0.5)	0.8(0.2)	0.0(0.0)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 8	%-OMIT
-- TOTAL --	18173	2761459(0%)	100.0(0.0) 260.7(0.5)	0.0
SEX				
MALE	9066	1380877(1%)	100.0(0.0) 257.0(0.6)	0.0
FEMALE	9106	1380477(1%)	100.0(0.0) 264.5(0.6)	0.0
ETHNICITY/RACE				
WHITE	12939	2044478(0%)	100.0(0.0) 266.7(0.6)	0.0
BLACK	2555	398198(1%)	100.0(0.0) 240.7(1.1)	0.0
HISPANIC	2043	241189(2%)	100.0(0.0) 242.4(1.3)	0.0
OTHER	636	77593(3%)	100.0(0.0) 263.6(1.6)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	1833	262530(5%)	100.0(0.0) 244.2(0.7)	0.0
GRADUATED H.S.	6444	974445(3%)	100.0(0.0) 255.5(0.7)	0.0
POST H.S.	8117	1261623(2%)	100.0(0.0) 271.8(0.7)	0.0
UNKNOWN	1609	234214(5%)	100.0(0.0) 241.6(1.1)	0.0
AGE				
13 YEARS OLD	12043	1888098(0%)	100.0(0.0) 266.5(0.6)	0.0
14 OR OLDER	5976	852015(1%)	100.0(0.0) 247.7(0.8)	0.0

Table 15(30)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	18172	2761355(0%)	50.0(0.5) 257.0(0.6)	50.0(0.5) 264.5(0.6)	0.0
SEX					
MALE	9066	1380877(1%)	100.0(0.0) 257.0(0.6)	0.0(0.0) *****(0.0)	0.0
FEMALE	9106	1380477(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 264.5(0.6)	0.0
ETHNICITY/RACE					
WHITE	12939	2044478(0%)	50.4(0.6) 263.0(0.7)	49.6(0.6) 270.4(0.8)	0.0
BLACK	2554	398095(1%)	48.4(1.2) 236.3(1.3)	51.6(1.2) 244.8(1.4)	0.0
HISPANIC	2043	241189(2%)	49.3(1.3) 237.6(2.0)	50.7(1.3) 247.0(1.6)	0.0
OTHER	636	77593(3%)	51.1(2.2) 259.3(3.1)	48.9(2.2) 268.1(2.7)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	1832	262426(5%)	44.0(1.7) 240.3(1.3)	56.0(1.7) 247.3(1.1)	0.0
GRADUATED H.S.	6444	974445(3%)	49.3(0.8) 251.0(0.8)	50.7(0.8) 260.0(0.8)	0.0
POST H.S.	8117	1261623(2%)	50.7(0.6) 267.8(0.9)	49.3(0.6) 275.9(0.8)	0.0
UNKNOWN	1609	234214(5%)	55.6(1.3) 241.2(1.6)	44.4(1.3) 242.2(1.2)	0.0
AGE					
13 YEARS OLD	12042	1887994(0%)	47.0(0.6) 263.3(0.7)	53.0(0.6) 269.3(0.7)	0.0
14 OR OLDER	5976	852015(1%)	57.0(0.7) 245.2(1.0)	43.0(0.7) 251.0(1.0)	0.0

Table 15(31)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	18173	2761459(0%)	74.0(0.2) 266.7(0.6)	14.4(0.1) 240.7(1.1)	8.7(0.1) 242.4(1.3)	1.1(0.1) 256.2(2.3)	1.6(0.2) 268.9(2.3)	0.0(0.0) 257.2(14.9)	0.0
SEX									
MALE	9066	1380877(1%)	74.6(0.4) 263.0(0.7)	13.9(0.4) 236.3(1.3)	8.6(0.2) 237.6(2.0)	1.3(0.1) 251.3(3.7)	1.6(0.1) 265.9(3.5)	0.0(0.0) 185.2(****)	0.0
FEMALE	9106	1380477(1%)	73.5(0.3) 270.4(0.8)	14.9(0.3) 244.8(1.4)	8.9(0.3) 247.0(1.6)	1.0(0.2) 262.3(2.9)	1.7(0.2) 271.6(4.0)	0.1(0.0) 264.9(13.1)	0.0
ETHNICITY/RACE									
WHITE	12939	2044478(0%)	100.0(0.0) 266.7(0.6)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	2555	398198(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 240.7(1.1)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	2043	241189(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 242.4(1.3)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	636	77593(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	40.3(5.0) 256.2(2.3)	58.4(5.1) 268.9(2.3)	1.3(0.5) 257.2(14.9)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1833	262530(5%)	58.2(2.2) 250.9(1.1)	18.6(1.3) 231.1(2.1)	20.8(1.9) 237.0(1.7)	1.6(0.4) 251.6(6.3)	0.8(0.2) 239.2(7.9)	0.0(0.0) 185.2(****)	0.0
GRADUATED H.S.	6444	974445(3%)	76.8(0.7) 259.7(0.8)	14.4(0.6) 239.9(1.5)	6.8(0.6) 242.6(2.1)	1.1(0.1) 248.0(3.6)	0.9(0.2) 257.4(6.3)	0.0(0.0) 235.0(9.4)	0.0
POST H.S.	8117	1261623(2%)	79.7(0.6) 276.1(0.8)	12.2(0.4) 249.2(1.9)	5.1(0.4) 256.6(1.8)	1.1(0.2) 266.3(3.9)	2.0(0.3) 278.1(3.2)	0.0(0.0) 290.5(9.5)	0.0
UNKNOWN	1609	234214(5%)	49.3(2.1) 250.8(1.3)	22.5(1.6) 228.6(1.8)	23.0(1.9) 231.5(1.7)	1.3(0.3) 245.7(9.0)	3.9(0.6) 259.4(3.5)	0.0(0.0) *****(0.0)	0.0
AGE									
13 YEARS OLD	12043	1888098(0%)	77.5(0.2) 270.9(0.6)	12.3(0.1) 247.7(1.4)	7.0(0.1) 249.3(1.7)	1.2(0.1) 261.8(3.0)	1.8(0.2) 273.7(2.9)	0.0(0.0) 264.9(13.1)	0.0
14 OR OLDER	5976	852015(1%)	66.4(0.4) 255.5(0.9)	18.9(0.4) 229.7(1.5)	12.5(0.4) 233.8(1.9)	1.0(0.1) 242.1(3.7)	1.1(0.1) 249.4(3.6)	0.0(0.0) 185.2(****)	0.0

Table 15(32)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	18173	2761459(0%)	22.7(0.4) 262.8(1.0)	23.5(1.5) 259.9(1.4)	26.3(1.3) 262.0(1.2)	27.5(0.6) 258.5(0.7)	0.0
SEX							
MALE	9066	1380877(1%)	23.5(0.6) 259.1(1.1)	23.3(1.4) 256.4(1.4)	26.0(1.4) 257.5(1.3)	27.2(0.9) 255.1(1.0)	0.0
FEMALE	9106	1380477(1%)	21.9(0.6) 266.7(1.4)	23.6(1.7) 263.4(1.7)	26.6(1.4) 266.4(1.4)	27.8(0.8) 261.8(0.8)	0.0
ETHNICITY/RACE							
WHITE	12939	2044478(0%)	24.0(0.1) 267.0(1.0)	21.5(1.7) 268.9(1.8)	30.1(1.6) 265.4(1.3)	24.4(0.1) 265.9(0.6)	0.0
BLACK	2555	398198(1%)	21.9(0.5) 244.3(3.9)	43.2(0.4) 239.1(1.3)	19.2(3.2) 240.2(1.9)	15.7(3.3) 240.7(2.3)	0.0
HISPANIC	2043	241189(2%)	14.6(3.8) 248.0(2.0)	10.8(4.1) 245.2(9.2)	8.5(2.3) 241.6(2.9)	66.1(0.8) 240.8(1.6)	0.0
OTHER	636	77593(3%)	19.3(4.9) 265.3(4.2)	12.9(2.9) 261.8(5.6)	19.6(3.6) 261.6(3.2)	48.2(7.0) 264.2(2.4)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1833	262530(5%)	18.9(2.3) 248.0(2.3)	31.0(2.2) 242.9(1.0)	21.0(2.6) 244.6(1.8)	29.1(3.0) 242.9(1.1)	0.0
GRADUATED H.S.	6444	974445(3%)	23.9(1.1) 258.4(1.1)	21.9(1.7) 251.8(1.3)	30.7(2.0) 257.6(1.4)	23.5(1.5) 253.5(1.4)	0.0
POST H.S.	8117	1261623(2%)	23.0(0.9) 272.3(1.2)	24.2(2.2) 273.3(2.0)	25.2(2.0) 272.0(0.9)	27.6(1.2) 269.8(1.1)	0.0
UNKNOWN	1609	234214(5%)	22.5(2.7) 245.8(1.3)	20.0(2.2) 240.1(2.5)	22.0(2.3) 244.6(2.5)	35.5(2.6) 238.0(1.9)	0.0
AGE							
13 YEARS OLD	12043	1888098(0%)	22.8(0.3) 269.1(1.0)	23.1(1.7) 267.8(1.5)	26.8(1.4) 265.6(1.3)	27.3(0.5) 264.1(0.6)	0.0
14 OR OLDER	5976	852015(1%)	21.8(0.7) 248.2(2.0)	24.4(1.4) 243.2(1.9)	25.5(1.5) 253.0(1.7)	28.3(1.1) 246.4(1.4)	0.0

Table 15(33)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT AGE

	N	WEIGHTED N	11-LESS	12	13	14	15	16-MORE	%-OMIT
-- TOTAL --	18173	2761459(0%)	0.0(0.0) 249.8(15.9)	0.8(0.1) 269.0(4.3)	68.4(0.2) 266.5(0.6)	26.1(0.2) 250.7(0.7)	4.0(0.2) 231.7(1.8)	0.8(0.1) 227.9(3.0)	0.0
SEX									
MALE	9066	1380877(1%)	0.0(0.0) *****(0.0)	0.6(0.1) 259.3(5.3)	64.2(0.4) 263.3(0.7)	29.3(0.4) 248.1(1.1)	5.0(0.3) 232.7(2.1)	0.9(0.2) 222.6(4.5)	0.0
FEMALE	9106	1380477(1%)	0.0(0.0) 249.8(15.9)	0.9(0.2) 275.5(4.7)	72.5(0.4) 269.3(0.7)	22.9(0.4) 254.1(1.1)	3.1(0.2) 230.2(3.0)	0.6(0.1) 236.2(5.1)	0.0
ETHNICITY/RACE									
WHITE	12939	2044478(0%)	0.0(0.0) 249.8(15.9)	0.7(0.1) 271.9(5.3)	71.6(0.2) 270.9(0.6)	24.7(0.3) 257.3(0.9)	2.6(0.2) 240.9(2.3)	0.3(0.1) 233.7(4.2)	0.0
BLACK	2555	398198(1%)	0.0(0.0) *****(0.0)	1.2(0.3) 265.3(9.1)	58.5(0.6) 247.7(1.4)	28.7(0.8) 232.4(1.7)	9.0(0.7) 222.5(2.5)	2.6(0.5) 226.1(4.2)	0.0
HISPANIC	2043	241189(2%)	0.0(0.0) *****(0.0)	0.5(0.2) 239.3(14.1)	55.2(0.9) 249.3(1.7)	35.1(1.1) 236.8(1.3)	7.8(1.1) 222.8(3.8)	1.3(0.6) 220.7(8.0)	0.0
OTHER	636	77593(3%)	0.0(0.0) *****(0.0)	1.3(0.4) 285.6(9.1)	74.9(1.1) 269.0(2.2)	19.8(1.2) 247.8(2.5)	3.5(0.8) 235.1(5.3)	0.5(0.3) 232.0(33.7)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1833	262530(5%)	0.0(0.0) *****(0.0)	0.6(0.2) 253.4(17.8)	52.9(1.2) 252.0(0.8)	34.3(1.1) 238.5(1.7)	10.0(1.0) 226.4(2.5)	2.2(0.6) 224.5(5.3)	0.0
GRADUATED H.S.	6444	974445(3%)	0.0(0.0) 249.8(15.9)	0.7(0.2) 259.8(7.1)	68.0(0.7) 260.4(0.7)	26.6(0.7) 246.6(1.2)	4.0(0.4) 235.6(3.1)	0.7(0.2) 229.1(5.1)	0.0
POST H.S.	8117	1261623(2%)	0.0(0.0) *****(0.0)	0.8(0.2) 278.7(4.8)	74.2(0.5) 275.2(0.7)	22.7(0.5) 263.4(1.2)	1.9(0.1) 240.6(3.0)	0.3(0.1) 234.3(5.7)	0.0
UNKNOWN	1609	234214(5%)	0.0(0.0) *****(0.0)	0.8(0.2) 260.0(8.2)	56.2(1.7) 250.0(1.3)	32.6(1.3) 232.9(1.7)	8.6(1.0) 220.7(3.9)	1.7(0.4) 230.0(5.8)	0.0
AGE									
13 YEARS OLD	12043	1688098(0%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 266.5(0.6)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
14 OR OLDER	5976	852015(1%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	84.5(0.9) 250.7(0.7)	13.1(0.6) 231.7(1.8)	2.5(0.4) 227.9(3.0)	0.0

Table 15(34)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	18173	2761459(0%)	5.3(1.1)	8.8(1.5)	10.6(2.3)	10.4(3.4)	16.5(3.0)	15.2(2.7)	33.1(2.3)	0.0
			259.9(2.3)	241.6(1.9)	276.9(2.6)	258.6(1.7)	262.1(1.2)	260.1(2.5)	261.1(0.9)	
SEX										
MALE	9066	1380877(1%)	5.5(1.1)	8.5(1.5)	10.5(2.4)	10.5(3.4)	16.9(3.1)	15.2(2.8)	32.9(2.4)	0.0
			255.2(3.1)	237.0(2.2)	273.5(2.5)	255.7(2.0)	258.4(1.2)	256.8(2.7)	256.9(1.0)	
FEMALE	9106	1380477(1%)	5.1(1.1)	9.2(1.6)	10.7(2.3)	10.3(3.4)	16.2(3.0)	15.3(2.6)	33.3(2.3)	0.0
			265.1(2.1)	245.9(2.2)	280.2(3.2)	261.5(2.0)	265.9(1.6)	263.3(2.6)	265.2(1.0)	
ETHNICITY/RACE										
WHITE	12939	2044478(0%)	6.0(1.2)	2.8(0.9)	12.6(2.9)	9.1(3.4)	18.1(3.3)	14.1(2.3)	37.3(2.6)	0.0
			264.3(2.1)	256.5(2.5)	277.8(2.9)	264.3(1.8)	264.4(1.2)	268.3(2.0)	265.1(0.9)	
BLACK	2555	398198(1%)	5.1(2.3)	30.4(4.7)	2.9(1.0)	12.8(4.4)	8.2(2.5)	16.3(2.9)	24.2(3.3)	0.0
			237.3(2.2)	237.2(2.1)	264.4(5.8)	244.8(2.1)	248.4(3.3)	241.8(2.2)	237.4(1.7)	
HISPANIC	2043	241189(2%)	1.0(0.4)	22.2(8.9)	5.0(1.5)	15.8(8.0)	15.6(5.9)	24.2(11.7)	16.3(3.7)	0.0
			240.9(7.3)	234.5(3.5)	263.4(2.3)	246.8(2.5)	247.3(3.3)	239.6(2.6)	241.8(2.6)	
OTHER	636	77593(3%)	2.7(0.9)	17.1(3.7)	13.6(3.4)	13.8(4.4)	20.5(4.4)	11.7(3.0)	20.6(3.8)	0.0
			245.0(5.8)	247.8(4.9)	283.4(2.4)	265.9(6.6)	271.5(4.0)	262.8(5.8)	257.1(3.2)	
PARENTAL EDUCATION										
NOT GRADUATED H.S.	1833	262530(5%)	6.3(1.8)	12.8(2.8)	1.6(0.5)	7.7(2.7)	12.7(3.6)	17.9(5.2)	41.0(2.9)	0.0
			246.2(2.8)	235.5(2.7)	251.9(6.2)	240.2(3.0)	244.2(2.1)	242.9(2.7)	247.7(1.2)	
GRADUATED H.S.	6444	974445(3%)	7.0(1.4)	8.4(1.4)	4.8(1.0)	10.2(3.2)	16.2(3.1)	13.8(2.7)	39.7(2.8)	0.0
			254.7(2.1)	241.5(2.6)	265.9(2.1)	254.6(2.0)	256.5(1.6)	256.0(1.3)	257.1(1.2)	
POST H.S.	8117	1261623(2%)	4.1(1.0)	6.5(1.1)	16.6(3.4)	11.2(4.0)	17.5(3.0)	15.6(2.5)	28.5(2.5)	0.0
			274.6(2.8)	251.2(2.4)	281.6(2.9)	266.7(1.6)	272.0(1.3)	271.8(1.5)	272.2(0.9)	
UNKNOWN	1609	234214(5%)	4.6(1.3)	19.2(4.4)	5.7(1.1)	10.7(3.3)	18.3(3.8)	17.0(4.5)	24.5(2.4)	0.0
			244.7(5.8)	230.9(3.3)	259.2(3.4)	244.4(2.1)	245.0(3.2)	237.3(4.9)	244.6(2.3)	
AGE										
13 YEARS OLD	12043	1888098(0%)	5.1(1.0)	7.9(1.4)	12.0(2.6)	10.7(3.6)	16.7(3.1)	14.4(2.4)	33.1(2.2)	0.0
			265.0(1.7)	248.1(2.7)	280.0(2.6)	263.2(1.6)	266.7(1.4)	267.3(2.0)	266.9(0.9)	
14 OR OLDER	5976	852015(1%)	5.7(1.3)	10.6(2.0)	7.3(1.7)	9.7(3.1)	16.0(3.1)	17.0(3.4)	33.5(2.5)	0.0
			249.7(3.7)	230.2(1.8)	265.5(3.4)	247.0(2.1)	251.1(2.1)	246.5(3.8)	248.3(1.3)	

Table 15(35)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	18003	2732812(1%)	9.6(0.5) 244.2(0.7)	35.7(1.0) 255.5(0.7)	46.2(1.1) 271.8(0.7)	8.6(0.4) 241.6(1.1)	1.0
SEX							
MALE	8975	1366061(2%)	8.4(0.6) 240.3(1.3)	35.2(1.0) 251.0(0.8)	46.8(1.1) 267.8(0.9)	9.5(0.5) 241.2(1.6)	1.1
FEMALE	9027	1366647(1%)	10.8(0.5) 247.3(1.1)	36.1(1.1) 260.0(0.8)	45.5(1.3) 275.9(0.8)	7.6(0.4) 242.2(1.2)	1.0
ETHNICITY/RACE							
WHITE	12812	2021783(1%)	7.6(0.5) 250.9(1.1)	37.0(1.1) 259.7(0.8)	49.7(1.4) 276.1(0.8)	5.7(0.3) 250.8(1.3)	1.1
BLACK	2537	395346(1%)	12.3(0.9) 231.1(2.1)	35.5(1.5) 239.9(1.5)	38.8(1.3) 249.2(1.9)	13.3(1.1) 228.6(1.8)	0.7
HISPANIC	2028	239352(2%)	22.8(2.1) 237.0(1.7)	27.9(2.7) 242.6(2.1)	26.8(2.1) 256.6(1.8)	22.5(2.2) 231.5(1.7)	0.8
OTHER	626	76331(3%)	8.2(1.5) 246.5(5.0)	25.4(2.4) 252.0(4.3)	50.4(2.6) 274.1(2.2)	15.9(1.8) 256.1(3.2)	1.6
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1833	262530(5%)	100.0(0.0) 244.2(0.7)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	6444	974445(3%)	0.0(0.0) *****(0.0)	100.0(0.0) 255.5(0.7)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	8117	1261623(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 271.8(0.7)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	1609	234214(5%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 241.6(1.1)	0.0
AGE							
13 YEARS OLD	11932	1868600(1%)	7.4(0.4) 252.0(0.8)	35.4(1.1) 260.4(0.7)	50.1(1.2) 275.2(0.7)	7.0(0.3) 250.0(1.3)	1.0
14 OR OLDER	5918	843042(1%)	14.5(0.8) 235.2(1.2)	36.2(1.2) 244.8(1.1)	37.4(1.2) 261.3(1.2)	11.9(0.8) 230.4(1.7)	1.1

Table 15(36)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	150	200	250	300	350
-- TOTAL --	18173	2761459(0%)	99.8(0.0)	95.5(0.2)	63.1(0.7)	12.5(0.4)	0.3(0.0)
SEX							
MALE	9066	1380877(1%)	99.8(0.1)	94.0(0.2)	59.1(0.8)	10.5(0.5)	0.1(0.0)
FEMALE	9106	1380477(1%)	99.9(0.0)	96.9(0.3)	67.1(0.8)	14.4(0.4)	0.5(0.1)
ETHNICITY/RACE							
WHITE	12939	2044478(0%)	99.9(0.0)	97.4(0.2)	69.9(0.8)	15.3(0.5)	0.4(0.1)
BLACK	2555	398198(1%)	99.4(0.2)	89.6(0.6)	40.0(1.5)	2.7(0.4)	0.0(0.0)
HISPANIC	2043	241189(2%)	99.6(0.1)	88.8(1.2)	42.0(1.7)	3.9(0.5)	0.1(0.1)
OTHER	636	77593(3%)	100.0(0.0)	95.5(0.7)	66.8(2.3)	13.9(1.9)	0.2(0.2)
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1833	262530(5%)	99.7(0.2)	91.8(0.5)	43.8(1.1)	4.0(0.4)	0.0(0.0)
GRADUATED H.S.	6444	974445(3%)	99.9(0.0)	95.1(0.4)	57.5(1.0)	8.0(0.4)	0.1(0.1)
POST H.S.	8117	1261623(2%)	99.9(0.0)	98.0(0.1)	75.4(0.8)	19.3(0.7)	0.6(0.1)
UNKNOWN	1609	234214(5%)	99.2(0.2)	88.3(0.8)	41.2(1.4)	4.3(0.5)	0.1(0.1)
AGE							
13 YEARS OLD	12043	1888098(0%)	100.0(0.0)	97.4(0.2)	69.9(0.8)	14.8(0.5)	0.5(0.1)
14 OR OLDER	5976	852015(1%)	99.5(0.1)	91.1(0.4)	47.7(1.0)	7.1(0.5)	0.0(0.0)

Table 15(37)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 11	%-OMIT
-- TOTAL --	19080	2563822(0%)	100.0(0.0) 289.3(0.8)	0.0
SEX				
MALE	9443	1292364(2%)	100.0(0.0) 284.5(1.0)	0.0
FEMALE	9637	1271457(2%)	100.0(0.0) 294.3(0.9)	0.0
ETHNICITY/RACE				
WHITE	13914	1906547(0%)	100.0(0.0) 295.8(0.9)	0.0
BLACK	2792	383493(1%)	100.0(0.0) 268.1(1.8)	0.0
HISPANIC	1699	203453(2%)	100.0(0.0) 269.5(2.0)	0.0
OTHER	675	70329(3%)	100.0(0.0) 287.6(2.2)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	2300	293458(5%)	100.0(0.0) 269.5(1.2)	0.0
GRADUATED H.S.	6600	865215(3%)	100.0(0.0) 281.8(0.7)	0.0
POST H.S.	9378	1301603(3%)	100.0(0.0) 300.6(0.9)	0.0
UNKNOWN	596	78893(5%)	100.0(0.0) 259.2(2.1)	0.0
AGE				
16 OR YOUNGER	1992	334011(6%)	100.0(0.0) 299.6(1.4)	0.0
17 YEARS OLD	14009	1699683(0%)	100.0(0.0) 295.0(0.7)	0.0
18 OR OLDER	3079	530128(3%)	100.0(0.0) 264.5(1.3)	0.0

Table 15(38)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	19080	2563822(0%)	50.4(0.8) 284.5(1.0)	49.6(0.8) 294.3(0.9)	0.0
SEX					
MALE	9443	1292364(2%)	100.0(0.0) 284.5(1.0)	0.0(0.0) *****(0.0)	0.0
FEMALE	9637	1271457(2%)	0.0(0.0) *****(0.0)	100.0(0.0) 294.3(0.9)	0.0
ETHNICITY/RACE					
WHITE	13914	1906547(0%)	50.3(0.8) 290.5(1.0)	49.7(0.8) 301.1(0.9)	0.0
BLACK	2792	383493(1%)	49.2(1.3) 264.3(2.2)	50.8(1.3) 271.8(1.9)	0.0
HISPANIC	1699	203453(2%)	51.3(1.7) 266.2(2.2)	48.7(1.7) 273.0(2.8)	0.0
OTHER	675	70329(3%)	56.9(2.9) 282.6(2.8)	43.1(2.9) 294.3(3.3)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	2300	293458(5%)	45.6(0.9) 264.5(1.5)	54.4(0.9) 273.7(1.6)	0.0
GRADUATED H.S.	6600	865215(3%)	50.0(0.8) 275.9(1.0)	50.0(0.8) 287.7(0.8)	0.0
POST H.S.	9378	1301603(3%)	51.2(1.2) 295.9(1.1)	48.8(1.2) 305.6(1.0)	0.0
UNKNOWN	596	78893(5%)	56.5(2.8) 258.4(2.8)	43.5(2.8) 260.3(3.3)	0.0
AGE					
16 OR YOUNGER	1992	334011(6%)	43.5(1.7) 295.2(1.9)	56.5(1.7) 303.1(1.6)	0.0
17 YEARS OLD	14009	1699683(0%)	48.2(0.8) 291.1(0.9)	51.8(0.8) 298.7(0.8)	0.0
18 OR OLDER	3079	530128(3%)	61.9(1.3) 263.3(1.3)	38.1(1.3) 266.6(1.7)	0.0

Table 15(39)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	19080	2563822(0%)	24.5(0.4) 290.9(2.7)	22.0(1.9) 287.3(1.7)	27.4(1.8) 290.7(1.7)	26.1(0.6) 288.2(0.8)	0.0
SEX							
MALE	9443	1292364(2%)	25.6(0.8) 286.3(2.4)	21.2(1.8) 282.7(2.6)	26.8(1.6) 285.8(1.7)	26.3(1.0) 282.8(1.5)	0.0
FEMALE	9637	1271457(2%)	23.4(0.8) 295.9(3.2)	22.8(2.1) 291.6(1.0)	28.0(2.1) 295.5(2.0)	25.8(1.1) 293.7(1.3)	0.0
ETHNICITY/RACE							
WHITE	13914	1906547(0%)	25.9(0.1) 296.0(2.8)	19.5(2.1) 297.8(1.7)	31.3(2.1) 294.7(1.3)	23.2(0.1) 295.2(1.3)	0.0
BLACK	2792	383493(1%)	24.5(0.5) 272.6(5.7)	42.4(0.5) 264.7(1.5)	17.8(3.5) 266.1(3.0)	15.3(3.7) 272.8(3.9)	0.0
HISPANIC	1699	203453(2%)	13.9(5.2) 264.0(2.2)	11.2(7.4) 273.5(14.5)	10.4(5.2) 265.0(5.6)	64.5(1.0) 270.7(2.2)	0.0
OTHER	675	70329(3%)	18.6(2.5) 284.5(5.7)	9.2(2.3) 294.1(3.9)	22.1(3.7) 280.5(8.3)	50.1(4.1) 290.8(2.4)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	2300	293458(5%)	21.8(2.7) 272.1(2.9)	28.1(2.4) 265.5(2.4)	22.3(3.0) 270.9(2.6)	27.7(2.8) 270.3(2.1)	0.0
GRADUATED H.S.	6600	865215(3%)	25.0(1.9) 282.5(1.4)	22.4(1.9) 278.9(1.8)	32.2(2.5) 284.6(1.2)	20.5(1.3) 279.7(1.4)	0.0
POST H.S.	9378	1301603(3%)	25.4(1.7) 301.7(3.1)	20.1(2.6) 301.3(1.6)	25.5(2.2) 301.5(1.3)	29.0(1.4) 298.5(1.1)	0.0
UNKNOWN	596	78893(5%)	21.5(2.8) 258.3(4.6)	20.4(3.2) 258.0(4.6)	21.8(4.0) 261.0(6.5)	36.2(2.7) 259.4(3.7)	0.0
AGE							
16 OR YOUNGER	1992	334011(6%)	33.5(2.7) 300.7(2.8)	26.3(3.7) 298.9(2.5)	17.8(3.4) 299.4(3.3)	22.5(2.1) 299.0(2.8)	0.0
17 YEARS OLD	14009	1699683(0%)	23.5(0.3) 295.5(2.2)	20.5(1.8) 294.5(1.6)	29.4(1.6) 296.2(1.3)	26.6(0.5) 293.8(0.8)	0.0
18 OR OLDER	3079	530128(3%)	22.0(1.8) 265.6(4.4)	24.2(2.2) 259.6(1.7)	27.1(3.2) 268.1(2.7)	26.7(1.7) 264.5(1.3)	0.0

Table 15(40)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	19080	2563822(0%)	74.4(0.2) 295.8(0.9)	15.0(0.1) 268.1(1.8)	7.9(0.2) 269.5(2.0)	0.8(0.1) 288.1(3.8)	1.9(0.1) 287.2(3.2)	0.0(0.0) 314.6(18.1)	0.0
SEX									
MALE	9443	1292364(2%)	74.2(0.4) 290.5(1.0)	14.6(0.4) 264.3(2.2)	8.1(0.3) 266.2(2.2)	1.0(0.1) 284.6(3.8)	2.1(0.1) 281.4(3.9)	0.0(0.0) 329.9(****)	0.0
FEMALE	9637	1271457(2%)	74.5(0.4) 301.1(0.9)	15.3(0.4) 271.8(1.9)	7.8(0.2) 273.0(2.8)	0.6(0.1) 294.2(5.6)	1.8(0.2) 294.1(4.1)	0.0(0.0) 308.2(20.2)	0.0
ETHNICITY/RACE									
WHITE	13914	1906547(0%)	100.0(0.0) 295.8(0.9)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	2792	383493(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 268.1(1.8)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	1699	203453(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 269.5(2.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	675	70329(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	28.5(2.5) 288.1(3.8)	70.9(2.5) 287.2(3.2)	0.7(0.3) 314.6(18.1)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	2300	293458(5%)	52.0(2.6) 277.6(1.6)	21.4(1.9) 259.0(2.6)	23.9(2.5) 261.8(1.6)	0.9(0.2) 259.4(9.9)	1.8(0.3) 268.3(6.1)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	6600	865215(3%)	75.5(0.8) 287.1(0.7)	16.6(0.6) 262.4(2.0)	6.0(0.6) 268.7(2.5)	0.9(0.1) 287.5(5.7)	1.0(0.2) 273.2(5.6)	0.0(0.0) 266.0(****)	0.0
POST H.S.	9378	1301603(3%)	80.3(0.7) 304.6(1.0)	11.9(0.6) 279.7(1.9)	4.7(0.6) 286.1(2.0)	0.7(0.1) 298.5(2.8)	2.3(0.1) 299.3(2.9)	0.0(0.0) 325.3(16.1)	0.0
UNKNOWN	596	78893(5%)	42.4(3.0) 272.2(3.1)	25.6(2.4) 249.8(4.3)	24.8(2.7) 247.2(2.2)	0.7(0.4) 253.5(19.3)	6.6(1.0) 258.5(5.5)	0.0(0.0) *****(0.0)	0.0
AGE									
16 OR YOUNGER	1992	334011(6%)	74.9(1.7) 305.1(1.5)	16.6(1.5) 279.6(2.8)	5.8(1.2) 283.5(3.4)	0.3(0.1) 287.7(11.6)	2.4(0.3) 308.1(4.2)	0.0(0.0) 329.9(****)	0.0
17 YEARS OLD	14009	1699683(0%)	78.7(0.2) 299.5(0.7)	12.1(0.2) 274.6(1.8)	6.6(0.2) 279.4(2.3)	0.9(0.1) 293.6(3.6)	1.7(0.1) 295.7(2.8)	0.0(0.0) 308.2(20.2)	0.0
18 OR OLDER	3079	530128(3%)	60.0(1.1) 272.8(1.7)	23.0(1.0) 252.1(2.1)	13.5(0.7) 250.2(1.9)	0.8(0.2) 269.7(9.5)	2.6(0.3) 257.9(5.5)	0.0(0.0) *****(0.0)	0.0

Table 15(41)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

IMPUTED STUDENT AGE

	N	WEIGHTED N	15-LESS	16	17	18	19	20-MORE	%-OMIT
-- TOTAL --	19080	2563822(0%)	0.2(0.0) 303.1(7.7)	12.8(0.8) 299.6(1.4)	66.3(0.2) 295.0(0.7)	17.8(0.6) 266.6(1.4)	2.4(0.2) 253.1(2.0)	0.5(0.1) 244.6(6.2)	0.0
SEX									
MALE	9443	1292364(2%)	0.1(0.0) 296.0(13.0)	11.1(0.8) 295.1(1.9)	63.3(0.5) 291.1(0.9)	21.8(0.7) 265.4(1.5)	3.1(0.3) 251.9(2.1)	0.5(0.1) 241.7(7.1)	0.0
FEMALE	9637	1271457(2%)	0.2(0.0) 306.8(8.9)	14.6(0.9) 303.0(1.6)	69.3(0.5) 298.7(0.8)	13.7(0.7) 268.6(1.8)	1.7(0.2) 255.4(3.8)	0.5(0.1) 248.1(9.3)	0.0
ETHNICITY/RACE									
WHITE	13914	1906547(0%)	0.2(0.0) 306.4(6.8)	13.0(0.8) 305.1(1.5)	70.2(0.2) 299.5(0.7)	15.3(0.7) 273.5(1.8)	1.2(0.1) 264.1(4.1)	0.2(0.0) 263.3(8.6)	0.0
BLACK	2792	383493(1%)	0.3(0.1) 301.5(24.4)	14.2(1.5) 279.1(2.9)	53.7(0.4) 274.6(1.8)	24.7(0.9) 254.2(2.3)	5.8(0.9) 246.7(3.4)	1.3(0.3) 234.0(8.0)	0.0
HISPANIC	1699	203453(2%)	0.1(0.1) 258.3(****)	9.4(2.2) 283.7(3.4)	55.2(1.2) 279.4(2.3)	27.4(1.1) 251.2(2.3)	6.2(0.8) 246.8(3.1)	1.7(0.4) 245.8(12.1)	0.0
OTHER	675	70329(3%)	0.3(0.0) 288.9(33.0)	12.5(1.1) 306.6(3.8)	61.4(1.3) 295.1(2.0)	18.1(1.8) 268.2(5.0)	5.6(1.4) 246.6(6.1)	2.1(0.4) 234.4(12.1)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	2300	293458(5%)	0.1(0.1) 262.4(15.3)	8.6(1.2) 285.0(3.1)	54.4(1.3) 279.1(1.2)	28.6(1.2) 254.1(2.0)	6.9(0.8) 245.4(3.3)	1.5(0.4) 237.3(14.3)	0.0
GRADUATED H.S.	6600	865215(3%)	0.2(0.1) 286.7(11.2)	10.9(0.9) 291.0(1.8)	66.9(0.7) 287.6(0.7)	19.4(0.8) 259.9(1.5)	2.2(0.3) 258.8(4.3)	0.4(0.1) 248.1(12.8)	0.0
POST H.S.	9378	1301603(3%)	0.2(0.1) 321.5(9.4)	15.3(0.8) 306.5(1.7)	69.4(0.5) 303.8(0.8)	13.7(0.7) 282.0(2.0)	1.1(0.1) 260.1(4.3)	0.3(0.1) 258.8(8.0)	0.0
UNKNOWN	596	78893(5%)	0.0(0.0) *****(0.0)	9.8(1.3) 269.2(6.5)	51.2(1.5) 269.2(2.9)	28.0(1.5) 244.2(3.2)	9.1(1.6) 244.9(5.3)	2.0(0.5) 230.2(13.4)	0.0
AGE									
16 OR YOUNGER	1992	334011(6%)	1.4(0.3) 303.1(7.7)	98.6(0.3) 299.6(1.4)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
17 YEARS OLD	14009	1699683(0%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 295.0(0.7)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
18 OR OLDER	3079	530128(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	86.0(1.0) 266.6(1.4)	11.6(0.8) 253.1(2.0)	2.5(0.3) 244.6(6.2)	0.0

Table 15(42)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	19080	2563822(0%)	5.3(1.2) 284.6(3.2)	10.6(2.3) 267.8(2.5)	16.4(2.7) 300.2(3.0)	8.8(2.1) 290.8(2.4)	10.2(2.7) 290.4(1.3)	16.8(1.5) 292.1(1.2)	31.9(1.6) 289.5(1.0)	0.0
SEX										
MALE	9443	1292364(2%)	5.4(1.2) 279.9(3.2)	10.3(2.2) 263.9(2.1)	17.6(2.6) 295.4(3.2)	7.5(2.1) 284.5(3.7)	10.2(2.7) 287.3(2.2)	16.9(2.0) 287.1(1.0)	32.1(1.6) 283.7(0.9)	0.0
FEMALE	9637	1271457(2%)	5.3(1.2) 289.5(3.5)	10.9(2.4) 271.5(3.0)	15.2(3.0) 305.9(2.9)	10.0(2.5) 295.6(1.7)	10.3(2.7) 293.6(1.7)	16.6(1.2) 297.3(1.8)	31.7(1.7) 295.5(1.4)	0.0
ETHNICITY/RACE										
WHITE	13914	1906547(0%)	5.2(1.1) 293.7(1.2)	4.0(1.4) 281.2(3.6)	18.5(3.1) 302.9(3.3)	7.2(2.0) 299.0(1.7)	11.0(2.9) 293.0(1.0)	17.4(1.1) 297.7(1.2)	36.8(1.9) 293.4(0.8)	0.0
BLACK	2792	383493(1%)	7.3(3.3) 259.4(3.0)	32.2(9.1) 262.1(2.9)	7.9(2.5) 283.0(5.4)	11.6(3.8) 273.4(2.6)	8.2(3.5) 279.2(4.7)	12.6(2.4) 270.9(2.5)	20.2(3.9) 265.8(2.3)	0.0
HISPANIC	1699	203453(2%)	4.6(3.2) 262.8(4.1)	28.6(9.3) 262.8(2.1)	12.5(6.9) 279.9(2.8)	15.6(6.5) 277.3(4.9)	6.1(2.9) 275.3(3.9)	19.7(8.2) 271.6(1.8)	13.0(4.9) 261.5(3.1)	0.0
OTHER	675	70329(3%)	2.2(0.9) 292.9(8.9)	20.7(5.2) 265.6(7.7)	17.1(4.3) 309.5(1.8)	15.3(5.1) 297.3(4.3)	12.9(4.4) 290.5(5.0)	14.3(2.1) 289.6(2.5)	17.5(2.2) 279.6(4.8)	0.0
PARENTAL EDUCATION										
NOT GRADUATED H.S.	2300	293458(5%)	8.4(2.5) 266.4(3.4)	18.5(4.4) 258.1(2.1)	3.8(0.9) 277.4(4.7)	8.6(2.4) 274.9(3.5)	7.6(2.5) 268.5(3.8)	14.4(2.6) 273.9(2.9)	38.6(2.2) 272.2(1.9)	0.0
GRADUATED H.S.	6600	865215(3%)	6.6(1.3) 282.1(3.0)	11.6(2.7) 265.0(3.1)	8.5(1.9) 281.5(2.7)	8.2(2.0) 285.9(2.6)	10.5(2.8) 284.1(1.8)	14.8(1.5) 283.2(1.6)	39.9(2.1) 284.7(1.0)	0.0
POST H.S.	9378	1301603(3%)	3.3(0.7) 301.0(2.5)	7.3(1.7) 280.8(2.6)	24.4(4.1) 306.1(2.3)	9.2(2.4) 298.9(2.1)	10.8(2.9) 298.9(1.6)	18.9(1.8) 301.3(1.7)	26.1(1.7) 302.0(1.0)	0.0
UNKNOWN	596	78893(5%)	7.0(2.0) 261.2(4.4)	26.6(5.6) 248.1(2.7)	9.7(3.0) 264.2(6.0)	9.7(2.8) 263.9(9.7)	10.5(3.3) 275.5(4.6)	15.6(2.5) 265.3(7.0)	20.9(2.3) 255.7(4.5)	0.0
AGE										
16 OR YOUNGER	1992	334011(6%)	4.5(1.1) 298.5(6.6)	10.8(2.9) 279.0(3.4)	21.1(4.1) 307.0(3.9)	10.3(2.9) 302.4(3.2)	11.4(3.0) 294.6(2.4)	17.2(3.0) 302.3(2.8)	24.6(2.5) 301.9(2.0)	0.0
17 YEARS OLD	14009	1699683(0%)	5.3(1.2) 290.2(2.7)	8.1(1.8) 276.5(2.7)	16.9(2.9) 303.5(2.6)	8.4(2.0) 29[illegible].1(1.9)	10.9(2.9) 295.2(1.2)	17.5(1.4) 296.6(1.4)	32.8(1.6) 294.9(0.9)	0.0
18 OR OLDER	3079	530128(3%)	6.0(1.4) 262.3(4.3)	18.5(3.6) 251.4(2.1)	11.7(2.0) 277.4(6.0)	9.0(2.4) 266.7(5.1)	7.3(2.0) 263.4(2.5)	14.0(2.2) 266.0(2.4)	33.4(2.3) 266.8(1.5)	0.0

Table 15(43)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	18874	2539169(1%)	11.6(0.6) 269.5(1.2)	34.1(1.0) 281.8(0.7)	51.3(1.3) 300.6(0.9)	3.1(0.2) 259.2(2.1)	1.0
SEX							
MALE	9319	1277103(2%)	10.5(0.6) 264.5(1.5)	33.9(1.1) 275.9(1.0)	52.2(1.4) 295.9(1.1)	3.5(0.2) 258.4(2.8)	1.2
FEMALE	9555	1262067(2%)	12.6(0.7) 273.7(1.6)	34.3(1.2) 287.7(0.8)	50.4(1.5) 305.6(1.0)	2.7(0.3) 260.3(3.3)	0.7
ETHNICITY/RACE							
WHITE	13736	1885096(1%)	8.1(0.6) 277.6(1.6)	34.7(1.3) 287.1(0.7)	55.5(1.6) 304.6(1.0)	1.8(0.1) 272.2(3.1)	1.1
BLACK	2776	381286(1%)	16.5(1.4) 259.0(2.6)	37.7(1.4) 262.4(2.0)	40.6(1.9) 279.7(1.9)	5.3(0.6) 249.8(4.3)	0.6
HISPANIC	1691	202723(2%)	34.5(4.4) 261.8(1.6)	25.7(2.5) 268.7(2.5)	30.1(3.6) 286.1(2.0)	9.6(1.2) 247.2(2.2)	0.4
OTHER	671	70065(3%)	11.3(1.3) 265.4(5.6)	22.9(1.9) 279.7(4.0)	57.5(2.2) 299.3(2.1)	8.2(1.2) 258.0(5.2)	0.4
PARENTAL EDUCATION							
NOT GRADUATED H.S.	2300	293458(5%)	100.0(0.0) 269.5(1.2)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	6600	865215(3%)	0.0(0.0) *****(0.0)	100.0(0.0) 281.8(0.7)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	9378	1301603(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 300.6(0.9)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	596	78893(5%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 259.2(2.1)	0.0
AGE							
16 OR YOUNGER	1974	331684(6%)	7.7(0.9) 284.8(3.1)	29.1(1.7) 290.9(1.8)	60.8(2.2) 306.7(1.7)	2.3(0.3) 269.2(6.5)	0.7
17 YEARS OLD	13849	1682103(1%)	9.5(0.6) 279.1(1.2)	34.4(1.1) 287.6(0.7)	53.7(1.3) 303.8(0.8)	2.4(0.1) 269.2(2.9)	1.0
18 OR OLDER	3051	525383(4%)	20.6(1.3) 251.8(1.7)	36.1(1.2) 259.6(1.5)	37.4(1.5) 280.0(1.9)	5.9(0.4) 243.7(2.8)	0.9

Table 15(44)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL READING PROFICIENCY MEANS - REPORTING VARIABLES
(MEANS ARE BASED ON A SINGLE SET OF PLAUSIBLE VALUES)

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	150	200	250	300	350
-- TOTAL --	19080	2563822(0%)	100.0(0.0)	98.7(0.1)	84.8(0.5)	40.2(0.8)	5.0(0.3)
SEX							
MALE	9443	1292364(2%)	100.0(0.0)	98.3(0.2)	81.1(0.8)	35.6(0.9)	3.7(0.3)
FEMALE	9637	1271457(2%)	100.0(0.0)	99.1(0.1)	88.5(0.5)	44.8(1.0)	6.2(0.4)
ETHNICITY/RACE							
WHITE	13914	1906547(0%)	100.0(0.0)	99.3(0.1)	89.6(0.5)	46.6(0.9)	6.1(0.3)
BLACK	2792	383493(1%)	100.0(0.0)	97.1(0.5)	68.8(1.5)	19.1(1.5)	1.0(0.2)
HISPANIC	1699	203453(2%)	100.0(0.0)	96.2(0.4)	71.0(2.0)	19.5(1.7)	1.4(0.4)
OTHER	675	70329(3%)	100.0(0.0)	99.0(0.4)	81.9(2.3)	40.0(1.9)	5.2(0.9)
PARENTAL EDUCATION							
NOT GRADUATED H.S.	2300	293458(5%)	100.0(0.0)	96.9(0.5)	70.8(1.3)	20.1(1.1)	1.1(0.3)
GRADUATED H.S.	6600	865215(3%)	100.0(0.0)	98.4(0.2)	81.7(0.8)	31.3(0.7)	2.6(0.2)
POST H.S.	9378	1301603(3%)	100.0(0.0)	99.5(0.1)	91.5(0.4)	52.2(0.9)	7.6(0.4)
UNKNOWN	596	78893(5%)	99.9(0.1)	94.3(1.0)	60.6(3.0)	13.6(2.2)	0.5(0.4)
AGE							
16 OR YOUNGER	1992	334011(6%)	100.0(0.0)	99.4(0.2)	92.2(0.6)	50.5(1.9)	7.8(1.0)
17 YEARS OLD	14009	1699683(0%)	100.0(0.0)	99.6(0.1)	89.2(0.5)	45.3(0.7)	5.6(0.3)
18 OR OLDER	3079	530128(3%)	100.0(0.0)	95.4(0.4)	66.1(1.1)	17.0(1.2)	1.1(0.2)

Table 15(45)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 4	%-OMIT
-- TOTAL --	8807	1408047(1%)	100.0(0.0) 1.58(0.01)	0.0
SEX				
MALE	4410	694799(2%)	100.0(0.0) 1.50(0.01)	0.0
FEMALE	4397	713248(1%)	100.0(0.0) 1.66(0.01)	0.0
ETHNICITY/RACE				
WHITE	5931	1016633(1%)	100.0(0.0) 1.63(0.01)	0.0
BLACK	1298	196446(2%)	100.0(0.0) 1.38(0.02)	0.0
HISPANIC	1169	152335(5%)	100.0(0.0) 1.46(0.02)	0.0
OTHER	409	42633(7%)	100.0(0.0) 1.60(0.03)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	526	77745(5%)	100.0(0.0) 1.43(0.03)	0.0
GRADUATED H.S.	1793	275493(4%)	100.0(0.0) 1.54(0.01)	0.0
POST H.S.	3364	549929(3%)	100.0(0.0) 1.66(0.01)	0.0
UNKNOWN	3067	495435(2%)	100.0(0.0) 1.53(0.01)	0.0
AGE				
9 YEARS OLD	5795	1026813(1%)	100.0(0.0) 1.60(0.01)	0.0
10 OR OLDER	2937	371667(2%)	100.0(0.0) 1.52(0.01)	0.0

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	8807	1408047(1%)	49.3(0.6) 1.50(0.01)	50.7(0.6) 1.66(0.01)	0.0
SEX					
MALE	4410	694799(2%)	100.0(0.0) 1.50(0.01)	0.0(0.0) *****(0.0)	0.0
FEMALE	4397	713248(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 1.66(0.01)	0.0
ETHNICITY/RACE					
WHITE	5931	1016633(1%)	49.2(0.8) 1.55(0.01)	50.8(0.8) 1.71(0.01)	0.0
BLACK	1298	196446(2%)	45.2(1.6) 1.30(0.02)	54.8(1.6) 1.45(0.02)	0.0
HISPANIC	1169	152335(5%)	54.1(1.7) 1.40(0.02)	45.9(1.7) 1.53(0.03)	0.0
OTHER	409	42633(7%)	53.9(2.6) 1.53(0.05)	46.1(2.6) 1.68(0.04)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	526	77745(5%)	46.4(3.4) 1.35(0.04)	53.6(3.4) 1.51(0.03)	0.0
GRADUATED H.S.	1793	275493(4%)	50.2(1.3) 1.45(0.01)	49.8(1.3) 1.64(0.02)	0.0
POST H.S.	3364	549929(3%)	51.9(0.9) 1.58(0.01)	48.1(0.9) 1.75(0.02)	0.0
UNKNOWN	3067	495435(2%)	46.5(1.2) 1.45(0.02)	53.5(1.2) 1.59(0.01)	0.0
AGE					
9 YEARS OLD	5795	1026813(1%)	46.7(0.7) 1.52(0.01)	53.3(0.7) 1.67(0.01)	0.0
10 OR OLDER	2937	371667(2%)	56.9(1.0) 1.45(0.02)	43.1(1.0) 1.60(0.02)	0.0

Table 15(47)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	8807	1408047(1%)	72.2(0.5) 1.63(0.01)	14.0(0.3) 1.38(0.02)	10.8(0.5) 1.46(0.02)	1.2(0.1) 1.56(0.04)	1.8(0.2) 1.63(0.04)	0.0(0.0) 1.80(0.25)	0.0
SEX									
MALE	4410	694799(2%)	72.0(0.8) 1.55(0.01)	12.8(0.6) 1.30(0.02)	11.9(0.5) 1.40(0.02)	1.6(0.2) 1.50(0.05)	1.7(0.3) 1.55(0.07)	0.0(0.0) 1.95(1.83)	0.0
FEMALE	4397	713248(1%)	72.4(0.8) 1.71(0.01)	15.1(0.5) 1.45(0.02)	9.8(0.6) 1.53(0.03)	0.9(0.1) 1.66(0.05)	1.8(0.2) 1.70(0.06)	0.0(0.0) 1.57(0.33)	0.0
ETHNICITY/RACE									
WHITE	5931	1016633(1%)	100.0(0.0) 1.63(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	1298	196446(2%)	0.0(0.0) *****(0.0)	100.0(0.0) 1.30(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	1169	152335(5%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.46(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	409	42633(7%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	40.7(4.4) 1.56(0.04)	58.1(4.4) 1.63(0.04)	1.1(0.7) 1.80(0.25)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	526	77745(5%)	64.5(3.1) 1.49(0.03)	14.2(2.1) 1.32(0.04)	19.1(2.1) 1.32(0.07)	1.5(0.5) 1.30(0.22)	0.7(0.3) 1.40(0.21)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	1793	275493(4%)	72.7(1.3) 1.60(0.02)	15.1(0.9) 1.38(0.02)	10.4(1.1) 1.42(0.03)	1.1(0.2) 1.54(0.08)	0.7(0.2) 1.41(0.09)	0.0(0.0) *****(0.0)	0.0
POST H.S.	3364	549929(3%)	73.7(0.8) 1.72(0.01)	13.5(0.9) 1.44(0.02)	9.4(0.6) 1.56(0.03)	1.3(0.2) 1.63(0.06)	2.1(0.3) 1.73(0.05)	0.0(0.0) 1.64(1.69)	0.0
UNKNOWN	3067	495435(2%)	72.0(0.9) 1.58(0.02)	13.6(0.7) 1.34(0.03)	11.1(0.7) 1.42(0.02)	1.1(0.2) 1.55(0.06)	2.1(0.3) 1.57(0.07)	0.1(0.1) 1.84(0.47)	0.0
AGE									
9 YEARS OLD	5795	1026813(1%)	73.9(0.6) 1.65(0.01)	12.6(0.3) 1.40(0.02)	10.5(0.5) 1.48(0.02)	1.1(0.1) 1.56(0.05)	1.8(0.2) 1.63(0.05)	0.0(0.0) 1.87(0.19)	0.0
10 OR OLDER	2937	371667(2%)	67.7(1.0) 1.58(0.01)	17.7(0.9) 1.35(0.03)	11.8(0.9) 1.41(0.03)	1.5(0.2) 1.55(0.08)	1.4(0.3) 1.56(0.07)	0.0(0.0) 1.50(1.31)	0.0

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	8807	1408047(1%)	22.5(0.5) 1.61(0.02)	23.6(1.4) 1.54(0.02)	27.3(1.3) 1.60(0.02)	26.6(0.7) 1.57(0.01)	0.0
SEX							
MALE	4410	694799(2%)	22.8(0.7) 1.53(0.02)	22.8(1.4) 1.47(0.02)	26.7(1.5) 1.51(0.02)	27.7(0.8) 1.49(0.02)	0.0
FEMALE	4397	713248(1%)	22.2(0.8) 1.68(0.02)	24.4(1.5) 1.61(0.02)	27.9(1.5) 1.68(0.02)	25.6(0.8) 1.65(0.01)	0.0
ETHNICITY/RACE							
WHITE	5931	1016633(1%)	23.2(0.4) 1.66(0.02)	21.1(1.6) 1.61(0.02)	31.4(1.6) 1.63(0.02)	24.3(0.5) 1.63(0.01)	0.0
BLACK	1298	196446(2%)	21.1(1.2) 1.39(0.04)	45.6(1.2) 1.38(0.03)	19.7(3.1) 1.40(0.04)	13.6(3.1) 1.37(0.03)	0.0
HISPANIC	1169	152335(5%)	20.8(4.0) 1.48(0.04)	14.3(4.3) 1.47(0.05)	12.0(2.8) 1.47(0.03)	52.9(2.3) 1.45(0.03)	0.0
OTHER	409	42633(7%)	17.4(4.1) 1.66(0.05)	14.1(3.0) 1.61(0.06)	19.6(3.4) 1.61(0.06)	48.9(6.1) 1.58(0.05)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	526	77745(5%)	18.3(1.5) 1.49(0.06)	32.0(2.8) 1.43(0.04)	23.5(2.9) 1.45(0.06)	26.3(2.3) 1.39(0.06)	0.0
GRADUATED H.S.	1793	275493(4%)	22.4(2.1) 1.57(0.03)	26.0(2.3) 1.51(0.03)	30.2(2.5) 1.56(0.02)	21.4(1.5) 1.53(0.03)	0.0
POST H.S.	3364	549929(3%)	21.9(1.3) 1.70(0.02)	22.5(1.6) 1.63(0.02)	27.1(1.8) 1.67(0.02)	28.5(1.4) 1.66(0.02)	0.0
UNKNOWN	3067	495435(2%)	23.9(1.2) 1.55(0.02)	22.1(2.0) 1.49(0.02)	26.8(1.9) 1.56(0.02)	27.2(1.2) 1.51(0.01)	0.0
AGE							
9 YEARS OLD	5795	1026813(1%)	22.9(0.6) 1.62(0.02)	23.4(1.5) 1.57(0.02)	27.3(1.5) 1.61(0.02)	26.4(0.6) 1.59(0.01)	0.0
10 OR OLDER	2937	371667(2%)	20.9(1.1) 1.56(0.04)	24.2(1.4) 1.46(0.02)	27.6(1.7) 1.55(0.03)	27.4(1.0) 1.51(0.02)	0.0

Table 15(49)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT AGE

	N	WEIGHTED N	7-LESS	8	9	10	11	12-MORE	%-OMIT
-- TOTAL --	8807	1408047(1%)	0.0(0.0) 1.72(1.68)	0.7(0.1) 1.55(0.07)	72.9(0.4) 1.60(0.01)	23.4(0.4) 1.53(0.01)	2.7(0.2) 1.40(0.04)	0.3(0.1) 1.35(0.07)	0.0
SEX									
MALE	4410	694799(2%)	0.0(0.0) *****(0.0)	0.5(0.1) 1.48(0.12)	69.0(0.8) 1.52(0.01)	26.6(0.8) 1.47(0.02)	3.5(0.3) 1.35(0.06)	0.4(0.1) 1.29(0.07)	0.0
FEMALE	4397	713248(1%)	0.0(0.0) 1.72(1.68)	0.8(0.2) 1.59(0.09)	76.7(0.5) 1.67(0.01)	20.3(0.5) 1.62(0.02)	2.0(0.2) 1.49(0.04)	0.1(0.1) 1.54(0.19)	0.0
ETHNICITY/RACE									
WHITE	5931	1016633(1%)	0.0(0.0) *****(0.0)	0.6(0.1) 1.61(0.08)	74.7(0.5) 1.65(0.01)	22.7(0.5) 1.59(0.01)	1.9(0.2) 1.49(0.04)	0.1(0.0) 1.44(0.12)	0.0
BLACK	1298	196446(2%)	0.0(0.0) *****(0.0)	0.9(0.4) 1.25(0.16)	65.7(1.3) 1.40(0.02)	26.6(1.4) 1.37(0.03)	6.1(0.7) 1.27(0.06)	0.7(0.2) 1.27(0.11)	0.0
HISPANIC	1169	152335(5%)	0.0(0.0) *****(0.0)	0.6(0.2) 1.44(0.14)	70.7(1.5) 1.48(0.02)	24.2(1.6) 1.42(0.02)	3.9(0.6) 1.34(0.08)	0.5(0.2) 1.32(0.08)	0.0
OTHER	409	42633(7%)	0.2(0.2) 1.72(1.68)	2.2(0.6) 1.81(0.14)	72.1(1.6) 1.61(0.03)	21.4(1.5) 1.56(0.06)	3.9(0.8) 1.52(0.12)	0.2(0.2) 1.48(0.47)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	526	77745(5%)	0.0(0.0) *****(0.0)	0.2(0.2) 0.97(0.24)	61.3(2.9) 1.45(0.03)	32.4(2.5) 1.41(0.04)	5.9(1.1) 1.40(0.09)	0.3(0.2) 1.33(0.12)	0.0
GRADUATED H.S.	1793	275493(4%)	0.0(0.0) *****(0.0)	0.4(0.1) 1.37(0.15)	69.1(1.2) 1.56(0.01)	26.9(1.1) 1.51(0.02)	3.4(0.4) 1.40(0.04)	0.2(0.1) 1.58(0.17)	0.0
POST H.S.	3364	549929(3%)	0.0(0.0) *****(0.0)	0.8(0.1) 1.66(0.12)	75.6(0.7) 1.68(0.01)	21.6(0.7) 1.62(0.02)	1.8(0.2) 1.42(0.09)	0.1(0.0) 1.36(0.16)	0.0
UNKNOWN	3067	495435(2%)	0.0(0.0) 1.72(1.68)	0.7(0.2) 1.47(0.09)	73.9(0.9) 1.54(0.02)	22.1(0.8) 1.50(0.02)	2.9(0.4) 1.40(0.06)	0.4(0.1) 1.35(0.13)	0.0
AGE									
9 YEARS OLD	5795	1026813(1%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.60(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
10 OR OLDER	2937	371667(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	88.7(0.7) 1.53(0.01)	10.3(0.7) 1.40(0.04)	1.0(0.2) 1.35(0.07)	0.0

Table 15(50)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	8807	1408047(1%)	6.2(0.9) 1.53(0.02)	11.6(2.0) 1.42(0.02)	14.8(2.2) 1.70(0.02)	7.4(1.6) 1.55(0.03)	11.1(1.6) 1.60(0.02)	16.6(1.3) 1.59(0.02)	32.5(1.6) 1.58(0.01)	0.0
SEX										
MALE	4410	694799(2%)	6.1(1.0) 1.42(0.03)	10.8(1.9) 1.34(0.02)	14.5(2.1) 1.61(0.03)	7.9(1.7) 1.48(0.04)	11.3(1.7) 1.52(0.02)	16.8(1.4) 1.52(0.02)	32.6(1.8) 1.50(0.01)	0.0
FEMALE	4397	713248(1%)	6.2(1.0) 1.63(0.03)	12.2(2.2) 1.50(0.03)	15.1(2.4) 1.79(0.03)	6.8(1.6) 1.63(0.04)	10.8(1.4) 1.68(0.04)	16.3(1.4) 1.65(0.03)	32.5(1.7) 1.66(0.02)	0.0
ETHNICITY/RACE										
WHITE	5931	1016633(1%)	6.6(1.1) 1.58(0.03)	6.0(1.6) 1.54(0.03)	16.3(2.5) 1.73(0.02)	5.9(1.6) 1.63(0.04)	12.1(1.6) 1.63(0.02)	16.8(1.3) 1.63(0.02)	36.4(1.9) 1.61(0.01)	0.0
BLACK	1298	196446(2%)	4.1(1.3) 1.34(0.03)	34.9(6.5) 1.33(0.03)	7.5(2.2) 1.55(0.09)	11.0(3.3) 1.37(0.03)	4.1(1.2) 1.44(0.09)	14.0(2.2) 1.43(0.07)	24.5(3.9) 1.39(0.02)	0.0
HISPANIC	1169	152335(5%)	6.9(2.5) 1.36(0.06)	19.5(5.2) 1.39(0.03)	11.7(2.4) 1.62(0.05)	11.6(3.2) 1.46(0.05)	11.0(3.3) 1.46(0.06)	18.4(4.2) 1.44(0.02)	20.8(3.5) 1.48(0.04)	0.0
OTHER	409	42633(7%)	3.9(2.2) 1.35(0.12)	7.8(1.7) 1.49(0.08)	23.5(4.0) 1.68(0.07)	10.5(2.6) 1.65(0.06)	17.9(5.4) 1.62(0.05)	16.4(2.8) 1.59(0.07)	19.9(3.3) 1.57(0.05)	0.0
PARENTAL EDUCATION										
NOT GRADUATED H.S.	526	77745(5%)	10.0(1.8) 1.35(0.10)	13.4(2.5) 1.29(0.05)	3.5(0.9) 1.50(0.13)	5.7(1.3) 1.43(0.08)	10.0(1.5) 1.50(0.10)	20.3(2.7) 1.46(0.04)	37.0(3.8) 1.46(0.04)	0.0
GRADUATED H.S.	1793	275493(4%)	9.2(1.4) 1.53(0.04)	11.0(2.3) 1.42(0.05)	7.2(1.4) 1.60(0.06)	5.2(1.3) 1.48(0.05)	9.8(1.4) 1.58(0.04)	14.8(1.6) 1.56(0.04)	42.8(2.7) 1.56(0.02)	0.0
POST H.S.	3364	549929(3%)	4.8(0.6) 1.61(0.04)	10.1(1.7) 1.50(0.03)	22.1(3.4) 1.76(0.02)	7.3(1.6) 1.63(0.04)	10.5(2.0) 1.66(0.03)	16.3(1.4) 1.65(0.03)	28.9(1.5) 1.67(0.02)	0.0
UNKNOWN	3067	495435(2%)	5.3(1.2) 1.50(0.05)	13.1(2.5) 1.38(0.03)	12.7(2.1) 1.63(0.03)	9.0(2.1) 1.52(0.04)	12.5(1.6) 1.56(0.03)	17.4(1.6) 1.55(0.03)	30.0(1.7) 1.52(0.01)	0.0
AGE										
9 YEARS OLD	5795	1026813(1%)	5.9(0.9) 1.54(0.04)	11.4(1.9) 1.44(0.02)	15.7(2.4) 1.72(0.02)	7.8(1.8) 1.57(0.03)	10.9(1.7) 1.62(0.02)	16.4(1.2) 1.61(0.03)	31.9(1.5) 1.60(0.01)	0.0
10 OR OLDER	2937	371667(2%)	7.1(1.1) 1.49(0.04)	11.7(2.3) 1.37(0.04)	11.9(1.8) 1.63(0.05)	5.8(1.2) 1.50(0.04)	11.5(1.6) 1.54(0.03)	17.0(1.9) 1.53(0.03)	34.9(2.2) 1.53(0.01)	0.0

Table 15(51)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	8750	1398603(1%)	5.6(0.3) 1.43(0.03)	19.7(0.8) 1.54(0.01)	39.3(1.0) 1.66(0.01)	35.4(0.8) 1.53(0.01)	0.7
SEX							
MALE	4381	690366(2%)	5.2(0.4) 1.35(0.04)	20.0(1.0) 1.45(0.01)	41.3(1.1) 1.58(0.01)	33.4(1.2) 1.45(0.02)	0.6
FEMALE	4369	708237(2%)	5.9(0.5) 1.51(0.03)	19.4(0.9) 1.64(0.02)	37.4(1.1) 1.75(0.02)	37.4(1.0) 1.59(0.01)	0.7
ETHNICITY/RACE							
WHITE	5910	1012419(1%)	5.0(0.4) 1.49(0.03)	19.8(0.9) 1.60(0.02)	40.0(1.2) 1.72(0.01)	35.2(1.0) 1.58(0.02)	0.4
BLACK	1282	194141(2%)	5.7(0.8) 1.32(0.04)	21.4(1.5) 1.38(0.02)	38.3(2.0) 1.44(0.02)	34.6(1.8) 1.34(0.03)	1.2
HISPANIC	1155	150310(5%)	9.9(1.3) 1.32(0.07)	19.1(1.8) 1.42(0.03)	34.5(1.7) 1.56(0.03)	36.6(2.0) 1.42(0.02)	1.3
OTHER	403	41733(6%)	4.0(1.0) 1.33(0.17)	11.8(1.6) 1.49(0.06)	44.7(3.0) 1.69(0.04)	39.5(3.2) 1.57(0.05)	2.1
PARENTAL EDUCATION							
NOT GRADUATED H.S.	526	77745(5%)	100.0(0.0) 1.43(0.03)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	1793	275493(4%)	0.0(0.0) *****(0.0)	100.0(0.0) 1.54(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	3364	549929(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.66(0.01)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	3067	495435(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.53(0.01)	0.0
AGE							
9 YEARS OLD	5758	1020049(1%)	4.7(0.3) 1.45(0.03)	18.7(0.9) 1.56(0.01)	40.8(1.0) 1.68(0.01)	35.9(0.9) 1.54(0.02)	0.7
10 OR OLDER	2917	368987(2%)	8.1(0.8) 1.41(0.04)	22.7(1.0) 1.50(0.02)	35.1(1.2) 1.60(0.02)	34.0(1.1) 1.48(0.02)	0.7

Table 15(52)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 4TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	1.0	2.0	3.0	4.0
-- TOTAL --	8807	1408047(1%)	92.4(0.4)	15.9(0.6)	0.0(0.0)	0.0(0.0)
SEX						
MALE	4410	694799(2%)	89.7(0.7)	10.6(0.6)	0.0(0.0)	0.0(0.0)
FEMALE	4397	713248(1%)	95.0(0.4)	21.2(0.9)	0.0(0.0)	0.0(0.0)
ETHNICITY/RACE						
WHITE	5931	1016633(1%)	94.6(0.5)	18.8(0.7)	0.0(0.0)	0.0(0.0)
BLACK	1298	196446(2%)	83.7(1.4)	6.1(0.9)	0.0(0.0)	0.0(0.0)
HISPANIC	1169	152335(5%)	88.7(0.9)	10.3(1.3)	0.0(0.0)	0.0(0.0)
OTHER	409	42633(7%)	92.7(1.8)	13.4(2.1)	0.0(0.0)	0.0(0.0)
PARENTAL EDUCATION						
NOT GRADUATED H.S.	526	77745(5%)	86.8(2.0)	8.4(0.9)	0.0(0.0)	0.0(0.0)
GRADUATED H.S.	1793	275493(4%)	91.7(0.6)	14.2(0.9)	0.0(0.0)	0.0(0.0)
POST H.S.	3364	549929(3%)	94.7(0.5)	21.2(0.8)	0.0(0.0)	0.0(0.0)
UNKNOWN	3067	495435(2%)	91.2(0.6)	12.5(0.8)	0.0(0.0)	0.0(0.0)
AGE						
9 YEARS OLD	5795	1026813(1%)	93.2(0.4)	17.4(0.6)	0.0(0.0)	0.0(0.0)
10 OR OLDER	2937	371667(2%)	90.1(0.9)	12.1(0.8)	0.0(0.0)	0.0(0.0)

Table 15(53)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 8	%-OMIT
-- TOTAL --	11092	1682192(1%)	100.0(0.0) 2.05(0.01)	0.0
SEX				
MALE	5486	839776(1%)	100.0(0.0) 1.96(0.01)	0.0
FEMALE	5606	842416(1%)	100.0(0.0) 2.14(0.01)	0.0
ETHNICITY/RACE				
WHITE	7916	1249837(1%)	100.0(0.0) 2.11(0.01)	0.0
BLACK	1500	232931(2%)	100.0(0.0) 1.86(0.01)	0.0
HISPANIC	1271	150212(4%)	100.0(0.0) 1.87(0.02)	0.0
OTHER	405	49212(4%)	100.0(0.0) 2.09(0.03)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	1072	152336(5%)	100.0(0.0) 1.89(0.02)	0.0
GRADUATED H.S.	3887	584033(3%)	100.0(0.0) 2.02(0.01)	0.0
POST H.S.	5038	785037(3%)	100.0(0.0) 2.13(0.01)	0.0
UNKNOWN	1000	144685(5%)	100.0(0.0) 1.90(0.02)	0.0
AGE				
13 YEARS OLD	7420	1155803(1%)	100.0(0.0) 2.08(0.01)	0.0
14 OR OLDER	3583	513626(2%)	100.0(0.0) 1.98(0.01)	0.0

Table 15(54)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	11092	1682192(1%)	49.9(0.6) 1.96(0.01)	50.1(0.6) 2.14(0.01)	0.0
SEX					
MALE	5486	839776(1%)	100.0(0.0) 1.96(0.01)	0.0(0.0) *****(0.0)	0.0
FEMALE	5606	842416(1%)	0.0(0.0) *****(0.0)	100.0(0.0) 2.14(0.01)	0.0
ETHNICITY/RACE					
WHITE	7916	1249837(1%)	50.3(0.7) 2.02(0.01)	49.7(0.7) 2.20(0.01)	0.0
BLACK	1500	232931(2%)	48.3(1.3) 1.78(0.02)	51.7(1.3) 1.94(0.02)	0.0
HISPANIC	1271	150212(4%)	48.8(1.5) 1.77(0.02)	51.2(1.5) 1.97(0.02)	0.0
OTHER	405	49212(4%)	49.9(2.6) 2.00(0.05)	50.1(2.6) 2.19(0.03)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	1072	152336(5%)	43.6(1.6) 1.79(0.03)	56.4(1.6) 1.97(0.02)	0.0
GRADUATED H.S.	3887	584033(3%)	49.5(0.8) 1.93(0.01)	50.5(0.8) 2.11(0.01)	0.0
POST H.S.	5038	785037(3%)	50.7(0.9) 2.04(0.01)	49.3(0.9) 2.23(0.01)	0.0
UNKNOWN	1000	144685(5%)	52.9(1.6) 1.84(0.03)	47.1(1.6) 1.98(0.03)	0.0
AGE					
13 YEARS OLD	7420	1155803(1%)	47.0(0.6) 1.99(0.01)	53.0(0.6) 2.16(0.01)	0.0
14 OR OLDER	3583	513626(2%)	56.5(1.2) 1.91(0.01)	43.5(1.2) 2.07(0.02)	0.0

Table 15(55)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	11092	1682192(1%)	74.3(0.4) 2.11(0.01)	13.8(0.2) 1.86(0.01)	8.9(0.3) 1.87(0.02)	1.1(0.1) 2.08(0.05)	1.8(0.2) 2.11(0.04)	0.0(0.0) 1.88(0.28)	0.0
SEX									
MALE	5486	839776(1%)	74.9(0.6) 2.02(0.01)	13.4(0.4) 1.78(0.02)	8.7(0.4) 1.77(0.02)	1.2(0.2) 1.97(0.08)	1.7(0.2) 2.02(0.06)	0.0(0.0) *****(0.0)	0.0
FEMALE	5606	842416(1%)	73.7(0.5) 2.20(0.01)	14.3(0.3) 1.94(0.02)	9.1(0.5) 1.97(0.02)	1.0(0.2) 2.20(0.05)	1.9(0.3) 2.18(0.05)	0.0(0.0) 1.88(0.28)	0.0
ETHNICITY/RACE									
WHITE	7916	1249837(1%)	100.0(0.0) 2.11(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	1500	232931(2%)	0.0(0.0) *****(0.0)	100.0(0.0) 1.86(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	1271	150212(4%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.87(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	405	49212(4%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	38.4(4.9) 2.08(0.05)	61.0(5.2) 2.11(0.04)	0.6(0.5) 1.88(0.28)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1072	152336(5%)	57.6(2.7) 1.95(0.02)	18.1(1.4) 1.76(0.05)	21.2(3.0) 1.84(0.04)	2.2(0.5) 1.92(0.11)	0.9(0.3) 1.95(0.07)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3887	584033(3%)	77.2(0.9) 2.06(0.01)	13.9(0.7) 1.87(0.02)	7.0(0.5) 1.89(0.03)	1.1(0.2) 2.05(0.07)	0.8(0.2) 2.00(0.10)	0.0(0.0) *****(0.0)	0.0
POST H.S.	5038	785037(3%)	79.2(0.6) 2.18(0.01)	12.1(0.5) 1.92(0.02)	5.5(0.3) 1.96(0.02)	0.9(0.2) 2.19(0.09)	2.2(0.3) 2.17(0.06)	0.0(0.0) 1.88(0.28)	0.0
UNKNOWN	1000	144685(5%)	52.5(2.6) 1.98(0.02)	19.2(1.9) 1.78(0.05)	23.1(2.6) 1.79(0.03)	1.1(0.2) 2.06(0.19)	4.2(0.8) 2.01(0.09)	0.0(0.0) *****(0.0)	0.0
AGE									
13 YEARS OLD	7420	1155803(1%)	77.7(0.3) 2.13(0.01)	12.1(0.3) 1.90(0.02)	7.2(0.2) 1.91(0.02)	1.1(0.2) 2.09(0.07)	2.0(0.2) 2.15(0.05)	0.0(0.0) 1.88(0.28)	0.0
14 OR OLDER	3583	513626(2%)	66.9(0.9) 2.05(0.01)	17.6(0.6) 1.81(0.02)	13.0(0.9) 1.83(0.02)	1.3(0.2) 2.04(0.05)	1.3(0.2) 1.95(0.06)	0.0(0.0) *****(0.0)	0.0

Table 15(56)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	11092	1682192(1%)	22.7(0.5) 2.09(0.01)	23.1(1.5) 2.03(0.02)	26.3(1.4) 2.06(0.01)	27.8(0.6) 2.03(0.02)	0.0
SEX							
MALE	5486	839776(1%)	23.3(0.7) 2.00(0.02)	22.9(1.6) 1.95(0.02)	26.3(1.5) 1.96(0.01)	27.5(0.7) 1.95(0.01)	0.0
FEMALE	5606	842416(1%)	22.1(0.6) 2.18(0.01)	23.3(1.6) 2.11(0.02)	26.4(1.5) 2.16(0.02)	28.2(0.9) 2.11(0.02)	0.0
ETHNICITY/RACE							
WHITE	7916	1249837(1%)	23.7(0.2) 2.13(0.01)	21.7(1.7) 2.11(0.02)	29.9(1.7) 2.09(0.01)	24.6(0.3) 2.10(0.02)	0.0
BLACK	1500	232931(2%)	22.6(0.8) 1.91(0.03)	41.2(0.8) 1.84(0.02)	20.8(3.5) 1.84(0.04)	15.4(3.4) 1.88(0.03)	0.0
HISPANIC	1271	150212(4%)	15.5(3.8) 1.91(0.03)	10.4(4.2) 1.91(0.05)	7.4(2.0) 1.81(0.06)	66.7(1.4) 1.87(0.02)	0.0
OTHER	405	49212(4%)	20.0(5.6) 2.14(0.06)	11.4(2.8) 2.06(0.06)	19.5(2.9) 2.10(0.05)	49.1(6.6) 2.08(0.06)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1072	152336(5%)	18.3(1.9) 1.94(0.04)	30.8(2.9) 1.86(0.03)	22.8(2.7) 1.92(0.04)	28.2(3.8) 1.86(0.04)	0.0
GRADUATED H.S.	3887	584033(3%)	24.6(1.4) 2.07(0.02)	21.8(1.7) 1.98(0.02)	30.2(2.1) 2.02(0.01)	23.5(1.6) 2.01(0.02)	0.0
POST H.S.	5038	785037(3%)	22.8(1.1) 2.16(0.01)	23.7(2.1) 2.13(0.03)	25.3(2.0) 2.13(0.02)	28.2(1.4) 2.11(0.02)	0.0
UNKNOWN	1000	144685(5%)	21.5(2.7) 1.90(0.03)	19.2(2.6) 1.91(0.05)	22.2(2.7) 1.95(0.03)	37.1(3.0) 1.87(0.02)	0.0
AGE							
13 YEARS OLD	7423	1155803(1%)	23.2(0.4) 2.11(0.01)	22.8(1.6) 2.08(0.02)	26.9(1.5) 2.08(0.02)	27.1(0.5) 2.07(0.02)	0.0
14 OR OLDER	3583	513626(2%)	20.8(1.3) 2.02(0.02)	24.1(1.6) 1.94(0.02)	25.4(1.6) 2.00(0.02)	29.7(1.3) 1.96(0.02)	0.0

Table 15(57)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT AGE

	N	WEIGHTED N	11-LESS	12	13	14	15	16-MORE	%-OMIT
-- TOTAL --	11092	1682192(1%)	0.0(0.0) 2.19(0.32)	0.7(0.1) 2.07(0.06)	68.7(0.5) 2.08(0.01)	26.0(0.5) 2.00(0.01)	3.9(0.2) 1.86(0.02)	0.7(0.1) 1.79(0.05)	0.0
SEX									
MALE	5486	839776(1%)	0.0(0.0) *****(0.0)	0.7(0.2) 1.98(0.10)	64.7(0.7) 1.99(0.01)	29.2(0.7) 1.93(0.01)	4.6(0.3) 1.81(0.03)	0.8(0.2) 1.69(0.09)	0.0
FEMALE	5606	842416(1%)	0.0(0.0) 2.19(0.32)	0.8(0.2) 2.15(0.06)	72.7(0.7) 2.16(0.01)	22.7(0.6) 2.09(0.02)	3.2(0.3) 1.92(0.05)	0.6(0.1) 1.91(0.09)	0.0
ETHNICITY/RACE									
WHITE	7916	1249837(1%)	0.0(0.0) 2.19(0.32)	0.7(0.2) 2.15(0.08)	71.8(0.4) 2.13(0.01)	24.6(0.5) 2.07(0.01)	2.6(0.1) 1.93(0.03)	0.3(0.1) 1.95(0.08)	0.0
BLACK	1500	232931(2%)	0.0(0.0) *****(0.0)	1.3(0.4) 1.90(0.08)	59.8(1.2) 1.90(0.02)	27.6(1.3) 1.82(0.02)	8.7(0.9) 1.79(0.06)	2.5(0.6) 1.72(0.05)	0.0
HISPANIC	1271	150212(4%)	0.0(0.0) *****(0.0)	0.5(0.2) 1.88(0.18)	55.1(1.9) 1.91(0.02)	35.9(1.2) 1.85(0.03)	7.3(1.7) 1.76(0.03)	1.2(0.7) 1.64(0.30)	0.0
OTHER	405	49212(4%)	0.0(0.0) *****(0.0)	0.8(0.3) 2.03(0.12)	72.9(2.1) 2.13(0.04)	23.0(1.9) 2.00(0.04)	2.9(0.7) 1.94(0.11)	0.4(0.3) 1.95(0.48)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1072	152336(5%)	0.1(0.1) 2.10(2.24)	0.7(0.4) 1.91(0.19)	52.7(1.6) 1.92(0.03)	34.1(1.2) 1.87(0.03)	10.7(1.1) 1.83(0.05)	1.7(0.5) 1.75(0.14)	0.0
GRADUATED H.S.	3887	584033(3%)	0.0(0.0) 2.27(0.30)	0.7(0.2) 2.13(0.07)	67.4(0.9) 2.05(0.01)	27.7(0.8) 1.98(0.01)	3.7(0.3) 1.86(0.05)	0.5(0.1) 1.97(0.08)	0.0
POST H.S.	5038	785037(3%)	0.0(0.0) *****(0.0)	0.8(0.2) 2.08(0.07)	74.9(0.8) 2.15(0.01)	22.1(0.7) 2.09(0.02)	1.8(0.2) 1.90(0.05)	0.4(0.1) 1.79(0.07)	0.0
UNKNOWN	1000	144685(5%)	0.0(0.0) *****(0.0)	0.8(0.3) 1.99(0.09)	57.6(2.2) 1.93(0.02)	30.9(1.8) 1.87(0.03)	8.8(1.0) 1.83(0.04)	1.8(0.5) 1.70(0.09)	0.0
AGE									
13 YEARS OLD	7420	1155803(1%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 2.08(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
14 OR OLDER	3583	513626(2%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	85.0(0.9) 2.00(0.01)	12.7(0.6) 1.86(0.02)	2.2(0.4) 1.79(0.05)	0.0

667

668

Table 15(58)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	11092	1682192(1%)	5.1(1.1) 2.03(0.03)	8.5(1.5) 1.88(0.02)	10.8(2.4) 2.21(0.02)	10.6(3.5) 2.01(0.02)	16.8(3.1) 2.07(0.02)	15.3(2.7) 2.04(0.03)	33.0(2.3) 2.05(0.01)	0.0
SEX										
MALE	5486	839776(1%)	5.0(1.0) 1.94(0.03)	7.8(1.5) 1.78(0.03)	11.1(2.5) 2.14(0.02)	10.8(3.6) 1.93(0.03)	17.5(3.2) 1.98(0.02)	15.3(2.9) 1.95(0.03)	32.4(2.3) 1.96(0.01)	0.0
FEMALE	5606	842416(1%)	5.2(1.1) 2.11(0.03)	9.2(1.6) 1.96(0.03)	10.5(2.3) 2.28(0.03)	10.3(3.5) 2.11(0.02)	16.1(3.0) 2.16(0.02)	15.2(2.7) 2.13(0.03)	33.5(2.4) 2.15(0.01)	0.0
ETHNICITY/RACE										
WHITE	7916	1249837(1%)	5.9(1.2) 2.06(0.02)	2.6(0.9) 2.04(0.03)	12.8(2.9) 2.22(0.02)	9.3(3.5) 2.08(0.02)	18.2(3.3) 2.10(0.02)	13.9(2.2) 2.11(0.02)	37.3(2.6) 2.09(0.01)	0.0
BLACK	1500	232931(2%)	4.2(1.9) 1.81(0.05)	30.2(4.9) 1.82(0.02)	2.9(1.0) 2.15(0.07)	13.4(4.8) 1.86(0.03)	8.9(2.7) 1.92(0.04)	16.5(3.1) 1.89(0.04)	24.0(3.4) 1.85(0.03)	0.0
HISPANIC	1271	150212(4%)	1.1(0.5) 1.88(0.12)	20.9(9.0) 1.81(0.03)	5.4(1.6) 2.02(0.08)	14.7(7.8) 1.90(0.03)	16.0(6.1) 1.91(0.04)	26.0(13.3) 1.86(0.04)	15.8(3.2) 1.87(0.04)	0.0
OTHER	405	49212(4%)	2.6(1.0) 2.01(0.11)	17.6(4.1) 1.99(0.08)	13.2(3.7) 2.23(0.04)	14.8(4.7) 1.99(0.08)	22.4(4.9) 2.15(0.07)	11.2(3.8) 2.04(0.13)	18.2(3.1) 2.14(0.06)	0.0
PARENTAL EDUCATION										
NOT GRADUATED H.S.	1072	152336(5%)	5.7(1.7) 1.88(0.07)	12.0(2.8) 1.81(0.06)	1.3(0.5) 2.04(0.11)	7.6(2.7) 1.85(0.06)	13.5(3.7) 1.91(0.03)	18.3(5.9) 1.89(0.04)	41.5(3.4) 1.91(0.03)	0.0
GRADUATED H.S.	3887	584033(3%)	6.8(1.4) 1.99(0.02)	8.3(1.5) 1.88(0.04)	4.9(1.1) 2.13(0.02)	10.2(3.3) 2.02(0.03)	16.3(3.1) 2.04(0.02)	13.7(2.7) 2.01(0.02)	39.8(2.8) 2.04(0.02)	0.0
POST H.S.	5038	785037(3%)	3.8(0.9) 2.14(0.03)	6.5(1.2) 1.95(0.03)	16.5(3.3) 2.25(0.02)	11.6(4.2) 2.05(0.02)	17.9(3.1) 2.14(0.02)	15.5(2.4) 2.12(0.03)	28.2(2.4) 2.14(0.02)	0.0
UNKNOWN	1000	144685(5%)	5.1(1.4) 1.95(0.06)	17.0(4.3) 1.79(0.06)	6.1(1.4) 2.12(0.05)	10.4(3.4) 1.90(0.06)	18.0(3.7) 1.91(0.04)	18.7(5.3) 1.93(0.04)	24.8(2.7) 1.89(0.04)	0.0
AGE										
13 YEARS OLD	7420	1155803(1%)	5.1(1.0) 2.06(0.03)	7.9(1.4) 1.92(0.03)	12.2(2.6) 2.22(0.02)	11.0(3.8) 2.05(0.02)	16.7(3.1) 2.09(0.02)	14.0(2.3) 2.07(0.02)	33.0(2.3) 2.09(0.01)	0.0
14 OR OLDER	3583	513626(2%)	5.2(1.3) 1.95(0.04)	9.6(2.1) 1.81(0.03)	7.4(1.9) 2.16(0.03)	9.5(3.3) 1.92(0.04)	16.8(3.3) 2.01(0.02)	18.2(3.9) 1.98(0.03)	33.3(2.7) 1.99(0.02)	0.0

Table 15(59)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	10997	1666092(1%)	9.1(0.4) 1.89(0.02)	35.1(1.1) 2.02(0.01)	47.1(1.2) 2.13(0.01)	8.7(0.5) 1.90(0.02)	1.0
SEX							
MALE	5429	830370(2%)	8.0(0.5) 1.79(0.03)	34.8(1.0) 1.93(0.01)	48.0(1.2) 2.04(0.01)	9.2(0.5) 1.84(0.03)	1.1
FEMALE	5568	835722(1%)	10.3(0.5) 1.97(0.02)	35.3(1.2) 2.11(0.01)	46.3(1.4) 2.23(0.01)	8.1(0.6) 1.98(0.03)	0.8
ETHNICITY/RACE							
WHITE	7842	1236555(1%)	7.1(0.4) 1.95(0.02)	36.5(1.2) 2.06(0.01)	50.3(1.4) 2.18(0.01)	6.1(0.4) 1.98(0.02)	1.1
BLACK	1492	231722(2%)	11.9(0.9) 1.76(0.05)	35.0(1.9) 1.87(0.02)	41.1(1.7) 1.92(0.02)	12.0(1.4) 1.78(0.05)	0.5
HISPANIC	1266	149566(4%)	21.6(2.7) 1.84(0.04)	27.3(3.1) 1.89(0.03)	28.8(2.5) 1.96(0.02)	22.3(2.5) 1.79(0.03)	0.4
OTHER	397	48248(4%)	9.8(2.2) 1.93(0.08)	23.2(3.0) 2.03(0.05)	51.3(3.5) 2.17(0.06)	15.7(2.6) 2.02(0.08)	2.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1072	152336(5%)	100.0(0.0) 1.89(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3887	584033(3%)	0.0(0.0) *****(0.0)	100.0(0.0) 2.02(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	5038	785037(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 2.13(0.01)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	1000	144685(5%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.90(0.02)	0.0
AGE							
13 YEARS OLD	7359	1144958(1%)	7.0(0.3) 1.92(0.03)	34.4(1.2) 2.05(0.01)	51.3(1.3) 2.15(0.01)	7.3(0.4) 1.93(0.02)	0.9
14 OR OLDER	3549	508371(2%)	13.9(0.8) 1.86(0.02)	36.7(1.1) 1.97(0.01)	37.5(1.4) 2.07(0.02)	11.8(1.0) 1.86(0.02)	1.0

Table 15(60)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 8TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	1.0	2.0	3.0	4.0
-- TOTAL --	11092	1682192(1%)	99.5(0.1)	55.0(0.6)	1.0(0.1)	0.0(0.0)
SEX						
MALE	5486	839776(1%)	99.1(0.1)	47.0(0.8)	0.5(0.1)	0.0(0.0)
FEMALE	5606	842416(1%)	99.9(0.1)	63.0(0.8)	1.4(0.1)	0.0(0.0)
ETHNICITY/RACE						
WHITE	7916	1249837(1%)	99.8(0.1)	60.6(0.7)	1.2(0.1)	0.0(0.0)
BLACK	1500	232931(2%)	98.7(0.3)	36.3(1.8)	0.1(0.0)	0.0(0.0)
HISPANIC	1271	150212(4%)	98.5(0.4)	36.8(1.4)	0.3(0.2)	0.0(0.0)
OTHER	405	49212(4%)	99.9(0.1)	57.2(4.0)	1.6(0.7)	0.0(0.0)
PARENTAL EDUCATION						
NOT GRADUATED H.S.	1072	152336(5%)	99.1(0.3)	40.5(1.9)	0.0(0.0)	0.0(0.0)
GRADUATED H.S.	3887	584033(3%)	99.6(0.1)	51.9(0.8)	0.4(0.1)	0.0(0.0)
POST H.S.	5038	785037(3%)	99.7(0.1)	63.1(0.9)	1.6(0.2)	0.0(0.0)
UNKNOWN	1000	144685(5%)	99.0(0.4)	39.5(1.5)	0.6(0.3)	0.0(0.0)
AGE						
13 YEARS OLD	7420	1155803(1%)	99.7(0.1)	57.6(0.7)	1.2(0.1)	0.0(0.0)
14 OR OLDER	3583	513626(2%)	99.1(0.1)	49.1(1.0)	0.5(0.1)	0.0(0.0)

Table 15(61)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT GRADE

	N	WEIGHTED N	GRADE 11	%-OMIT
-- TOTAL --	10657	1430241(1%)	100.0(0.0) 2.19(0.01)	0.0
SEX				
MALE	5215	714418(2%)	100.0(0.0) 2.09(0.01)	0.0
FEMALE	5442	715823(2%)	100.0(0.0) 2.29(0.01)	0.0
ETHNICITY/RACE				
WHITE	7892	1077899(1%)	100.0(0.0) 2.24(0.01)	0.0
BLACK	1478	205670(2%)	100.0(0.0) 2.00(0.02)	0.0
HISPANIC	902	107250(3%)	100.0(0.0) 2.00(0.02)	0.0
OTHER	385	39422(4%)	100.0(0.0) 2.16(0.03)	0.0
PARENTAL EDUCATION				
NOT GRADUATED H.S.	1267	159736(5%)	100.0(0.0) 1.99(0.02)	0.0
GRADUATED H.S.	3675	479173(3%)	100.0(0.0) 2.15(0.01)	0.0
POST H.S.	5312	740485(3%)	100.0(0.0) 2.27(0.01)	0.0
UNKNOWN	290	37087(7%)	100.0(0.0) 1.99(0.03)	0.0
AGE				
16 OR YOUNGER	1102	184232(6%)	100.0(0.0) 2.23(0.02)	0.0
17 YEARS OLD	7919	963071(1%)	100.0(0.0) 2.21(0.01)	0.0
18 OR OLDER	1636	282938(3%)	100.0(0.0) 2.08(0.02)	0.0

Table 15(62)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

STUDENT SEX

	N	WEIGHTED N	MALE	FEMALE	%-OMIT
-- TOTAL --	10657	1430241(1%)	50.0(0.9) 2.09(0.01)	50.0(0.9) 2.29(0.01)	0.0
SEX					
MALE	5215	714418(2%)	100.0(0.0) 2.09(0.01)	0.0(0.0) *****(0.0)	0.0
FEMALE	5442	715823(2%)	0.0(0.0) *****(0.0)	100.0(0.0) 2.29(0.01)	0.0
ETHNICITY/RACE					
WHITE	7892	1077899(1%)	49.6(0.9) 2.14(0.01)	50.4(0.9) 2.35(0.01)	0.0
BLACK	1478	205670(2%)	49.2(1.9) 1.91(0.02)	50.8(1.9) 2.09(0.02)	0.0
HISPANIC	902	107250(3%)	52.1(2.0) 1.92(0.03)	47.9(2.0) 2.09(0.03)	0.0
OTHER	385	39422(4%)	56.5(3.0) 2.09(0.04)	43.5(3.0) 2.26(0.06)	0.0
PARENTAL EDUCATION					
NOT GRADUATED H.S.	1267	159736(5%)	46.6(1.4) 1.90(0.03)	53.4(1.4) 2.06(0.03)	0.0
GRADUATED H.S.	3675	479173(3%)	50.0(0.9) 2.05(0.01)	50.0(0.9) 2.25(0.02)	0.0
POST H.S.	5312	740485(3%)	50.1(1.3) 2.16(0.01)	49.9(1.3) 2.37(0.01)	0.0
UNKNOWN	290	37087(7%)	58.1(3.3) 1.93(0.04)	41.9(3.3) 2.07(0.06)	0.0
AGE					
16 OR YOUNGER	1102	184232(6%)	43.2(2.2) 2.11(0.03)	56.8(2.2) 2.32(0.02)	0.0
17 YEARS OLD	7919	963071(1%)	47.7(0.9) 2.11(0.01)	52.3(0.9) 2.30(0.01)	0.0
18 OR OLDER	1636	282938(3%)	62.1(1.8) 2.02(0.02)	37.9(1.8) 2.18(0.03)	0.0

Table 15(63)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

ETHNICITY/RACE

	N	WEIGHTED N	WHITE	BLACK	HISPANIC	AMER IND	ASIAN	UNCLASS	%-OMIT
-- TOTAL --	10657	1430241(1%)	75.4(0.3) 2.24(0.01)	14.4(0.2) 2.00(0.02)	7.5(0.2) 2.00(0.02)	0.9(0.1) 2.10(0.05)	1.9(0.1) 2.19(0.04)	0.0(0.0) 2.04(0.34)	0.0
SEX									
MALE	5215	714418(2%)	74.9(0.6) 2.14(0.01)	14.2(0.6) 1.91(0.02)	7.8(0.3) 1.92(0.03)	1.1(0.1) 2.02(0.05)	2.0(0.2) 2.13(0.07)	0.0(0.0) *****(0.0)	0.0
FEMALE	5442	715823(2%)	75.8(0.6) 2.35(0.01)	14.6(0.5) 2.09(0.02)	7.2(0.3) 2.09(0.03)	0.6(0.1) 2.25(0.09)	1.8(0.2) 2.27(0.07)	0.0(0.0) 2.04(0.34)	0.0
ETHNICITY/RACE									
WHITE	7892	1077899(1%)	100.0(0.0) 2.24(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
BLACK	1478	205670(2%)	0.0(0.0) *****(0.0)	100.0(0.0) 2.00(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
HISPANIC	902	107250(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 2.00(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
OTHER	385	39422(4%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	31.5(3.2) 2.10(0.05)	68.0(3.2) 2.19(0.04)	0.5(0.3) 2.04(0.34)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1267	159736(5%)	53.9(2.7) 2.07(0.03)	20.1(1.8) 1.85(0.04)	24.0(2.5) 1.91(0.04)	0.6(0.2) 1.85(0.15)	1.4(0.3) 2.06(0.16)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3675	479173(3%)	77.0(0.9) 2.19(0.01)	15.5(0.6) 1.98(0.03)	5.5(0.7) 2.02(0.03)	1.0(0.1) 2.05(0.11)	1.0(0.2) 2.07(0.08)	0.0(0.0) *****(0.0)	0.0
POST H.S.	5312	740485(3%)	80.2(0.8) 2.30(0.01)	11.9(0.7) 2.08(0.02)	4.6(0.5) 2.10(0.03)	0.9(0.1) 2.18(0.06)	2.4(0.2) 2.25(0.05)	0.0(0.0) 2.04(0.34)	0.0
UNKNOWN	290	37087(7%)	44.8(4.4) 2.05(0.05)	25.5(2.8) 1.94(0.06)	22.4(3.3) 1.89(0.07)	1.3(0.7) 2.01(0.24)	6.0(1.5) 2.07(0.13)	0.0(0.0) *****(0.0)	0.0
AGE									
16 OR YOUNGER	1102	184232(6%)	74.9(1.9) 2.28(0.02)	16.4(1.9) 2.06(0.04)	5.8(1.3) 2.09(0.06)	0.4(0.2) 2.10(0.13)	2.5(0.4) 2.23(0.10)	0.0(0.0) *****(0.0)	0.0
17 YEARS OLD	7919	963071(1%)	79.4(0.3) 2.25(0.01)	11.7(0.2) 2.02(0.03)	6.4(0.2) 2.02(0.03)	0.9(0.1) 2.11(0.07)	1.5(0.1) 2.23(0.04)	0.0(0.0) 2.04(0.34)	0.0
18 OR OLDER	1636	282938(3%)	61.8(1.5) 2.16(0.02)	22.3(1.0) 1.93(0.03)	12.2(0.8) 1.93(0.03)	1.0(0.2) 2.06(0.09)	2.6(0.3) 2.09(0.07)	0.0(0.0) *****(0.0)	0.0

Table 15(64)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

REGION

	N	WEIGHTED N	NE	SE	CENTRAL	WEST	%-OMIT
-- TOTAL --	10657	1430241(1%)	24.8(0.5) 2.22(0.03)	21.9(1.9) 2.16(0.02)	27.3(1.8) 2.20(0.02)	26.0(0.7) 2.17(0.01)	0.0
SEX							
MALE	5215	714418(2%)	25.5(0.7) 2.12(0.02)	21.2(2.0) 2.05(0.02)	26.5(1.6) 2.10(0.02)	26.8(1.2) 2.07(0.01)	0.0
FEMALE	5442	715823(2%)	24.1(0.9) 2.32(0.04)	22.7(2.1) 2.25(0.02)	28.2(2.2) 2.30(0.03)	25.1(1.3) 2.27(0.02)	0.0
ETHNICITY/RACE							
WHITE	7892	1077899(1%)	26.2(0.3) 2.26(0.03)	19.9(2.2) 2.23(0.02)	30.9(2.1) 2.23(0.02)	23.0(0.3) 2.24(0.01)	0.0
BLACK	1478	205670(2%)	24.6(0.7) 2.04(0.04)	40.8(0.9) 1.98(0.03)	18.1(3.3) 2.00(0.03)	16.5(3.7) 2.00(0.03)	0.0
HISPANIC	902	107250(3%)	13.4(5.0) 1.95(0.03)	11.0(7.3) 2.03(0.08)	10.2(4.7) 1.97(0.11)	65.4(1.3) 2.01(0.02)	0.0
OTHER	385	39422(4%)	17.4(3.3) 2.15(0.08)	10.3(2.8) 2.20(0.10)	22.9(3.7) 2.14(0.08)	49.4(5.0) 2.17(0.05)	0.0
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1267	159736(5%)	22.4(2.7) 2.03(0.07)	28.2(2.5) 1.95(0.03)	22.0(3.0) 2.02(0.04)	27.4(2.8) 1.95(0.04)	0.0
GRADUATED H.S.	3675	479173(3%)	25.5(2.0) 2.17(0.02)	21.8(2.1) 2.11(0.02)	32.1(2.6) 2.17(0.02)	20.6(1.5) 2.13(0.02)	0.0
POST H.S.	5312	740485(3%)	25.3(1.8) 2.29(0.03)	20.2(2.7) 2.25(0.02)	25.4(2.2) 2.28(0.02)	29.1(1.5) 2.24(0.01)	0.0
UNKNOWN	290	37087(7%)	21.8(3.8) 2.02(0.05)	22.8(4.0) 1.95(0.07)	24.0(4.8) 2.00(0.10)	31.4(2.9) 1.98(0.05)	0.0
AGE							
16 OR YOUNGER	1102	184232(6%)	32.1(2.8) 2.25(0.04)	26.2(3.7) 2.21(0.04)	18.6(3.2) 2.24(0.05)	23.1(2.1) 2.22(0.04)	0.0
17 YEARS OLD	7919	963071(1%)	24.3(0.4) 2.24(0.03)	20.5(1.9) 2.19(0.02)	29.1(1.7) 2.22(0.02)	26.2(0.7) 2.19(0.01)	0.0
18 OR OLDER	1636	282938(3%)	21.8(1.9) 2.11(0.03)	24.0(2.1) 2.02(0.03)	27.1(3.2) 2.12(0.05)	27.1(1.7) 2.07(0.02)	0.0

Table 15(65)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

IMPUTED STUDENT AGE

	N	WEIGHTED N	15-LESS	16	17	18	19	20-MORE	%-OMIT
-- TOTAL --	10657	1430241(1%)	0.2(0.0) 2.30(0.14)	12.7(0.7) 2.23(0.02)	67.3(0.4) 2.21(0.01)	17.3(0.5) 2.10(0.02)	2.1(0.2) 1.94(0.04)	0.4(0.1) 1.99(0.11)	0.0
SEX									
MALE	5215	714418(2%)	0.1(0.1) 2.16(0.16)	11.0(0.8) 2.11(0.03)	64.2(0.7) 2.11(0.01)	21.4(0.7) 2.04(0.02)	2.7(0.3) 1.89(0.05)	0.5(0.1) 1.94(0.12)	0.0
FEMALE	5442	715823(2%)	0.2(0.1) 2.39(0.17)	14.4(0.9) 2.32(0.03)	70.4(0.8) 2.30(0.01)	13.2(0.8) 2.20(0.03)	1.4(0.2) 2.04(0.07)	0.3(0.1) 2.06(0.15)	0.0
ETHNICITY/RACE									
WHITE	7892	1077899(1%)	0.1(0.0) 2.34(0.15)	12.7(0.8) 2.28(0.02)	71.0(0.4) 2.25(0.01)	14.9(0.7) 2.17(0.02)	1.1(0.1) 2.01(0.08)	0.2(0.1) 2.23(0.20)	0.0
BLACK	1478	205670(2%)	0.2(0.1) 2.20(0.22)	14.5(1.7) 2.06(0.04)	54.6(1.0) 2.02(0.03)	24.0(1.1) 1.95(0.03)	5.7(1.0) 1.87(0.06)	1.0(0.4) 1.79(0.15)	0.0
HISPANIC	902	107250(3%)	0.3(0.3) 2.24(1.79)	9.8(2.1) 2.08(0.05)	57.7(1.8) 2.02(0.03)	26.4(1.8) 1.94(0.04)	4.5(1.1) 1.88(0.10)	1.4(0.4) 1.90(0.17)	0.0
OTHER	385	39422(4%)	0.9(0.4) 2.29(0.14)	12.6(1.6) 2.20(0.09)	60.6(2.1) 2.19(0.03)	20.9(1.7) 2.07(0.06)	3.2(0.8) 2.13(0.19)	1.7(0.7) 2.07(0.26)	0.0
PARENTAL EDUCATION									
NOT GRADUATED H.S.	1267	159736(5%)	0.2(0.1) 2.13(0.42)	7.7(1.2) 2.07(0.08)	55.6(2.1) 2.02(0.02)	29.4(2.0) 1.94(0.03)	6.2(1.0) 1.83(0.10)	0.9(0.3) 1.82(0.16)	0.0
GRADUATED H.S.	3675	479173(3%)	0.2(0.1) 2.35(0.15)	11.4(1.0) 2.15(0.03)	68.4(0.9) 2.17(0.01)	17.9(0.8) 2.09(0.02)	1.8(0.3) 1.99(0.09)	0.3(0.1) 2.09(0.21)	0.0
POST H.S.	5312	740485(3%)	0.2(0.1) 2.30(0.20)	14.8(0.8) 2.29(0.02)	69.8(0.7) 2.28(0.01)	13.7(0.7) 2.20(0.03)	1.2(0.2) 2.02(0.08)	0.3(0.1) 2.03(0.25)	0.0
UNKNOWN	290	37087(7%)	0.0(0.0) *****(0.0)	9.7(2.0) 2.07(0.15)	54.6(3.0) 1.99(0.04)	27.7(2.9) 1.96(0.10)	5.7(1.6) 1.96(0.10)	2.3(0.9) 1.96(0.23)	0.0
AGE									
16 OR YOUNGER	1102	184232(6%)	1.4(0.4) 2.30(0.14)	98.6(0.4) 2.23(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
17 YEARS OLD	7919	963071(1%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 2.21(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
18 OR OLDER	1636	282938(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	87.3(1.0) 2.10(0.02)	10.5(0.9) 1.94(0.04)	2.2(0.3) 1.99(0.11)	0.0

Table 15(66)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

SIZE/TYPE OF COMMUNITY

	N	WEIGHTED N	RURAL	DIS URB	ADV URB	BIG CITY	FRINGE	MEDIUM	SMALL	%-OMIT
-- TOTAL --	10657	1430241(1%)	5.5(1.2) 2.13(0.03)	10.0(2.2) 2.01(0.02)	16.8(2.7) 2.28(0.02)	8.8(2.1) 2.18(0.02)	9.9(2.7) 2.19(0.02)	16.9(1.5) 2.21(0.02)	32.1(1.6) 2.19(0.01)	0.0
SEX										
MALE	5215	714418(2%)	5.7(1.2) 2.03(0.04)	9.2(2.0) 1.92(0.02)	18.6(2.6) 2.19(0.02)	7.4(2.1) 2.05(0.03)	9.7(2.7) 2.07(0.03)	16.7(2.1) 2.10(0.02)	32.7(1.8) 2.09(0.02)	0.0
FEMALE	5442	715823(2%)	5.4(1.2) 2.23(0.04)	10.7(2.5) 2.09(0.02)	15.0(2.9) 2.40(0.03)	10.2(2.5) 2.27(0.02)	10.1(2.7) 2.30(0.03)	17.1(1.2) 2.31(0.03)	31.5(1.8) 2.30(0.02)	0.0
ETHNICITY/RACE										
WHITE	7892	1077899(1%)	5.3(1.1) 2.20(0.03)	4.0(1.5) 2.13(0.03)	18.6(3.0) 2.31(0.02)	7.2(2.0) 2.25(0.02)	10.4(2.8) 2.22(0.03)	17.5(1.2) 2.26(0.02)	36.9(2.0) 2.22(0.02)	0.0
BLACK	1478	205670(2%)	7.5(3.5) 1.94(0.04)	30.6(8.9) 1.97(0.02)	8.8(3.1) 2.13(0.06)	12.3(3.9) 2.03(0.03)	8.7(3.7) 2.07(0.06)	12.9(2.6) 2.00(0.05)	19.2(3.6) 1.97(0.04)	0.0
HISPANIC	902	107250(3%)	4.8(3.0) 1.99(0.15)	27.2(8.9) 1.90(0.03)	12.5(6.5) 2.05(0.05)	16.0(6.7) 2.06(0.04)	7.0(3.3) 2.03(0.05)	19.1(8.5) 2.03(0.03)	13.4(4.8) 2.01(0.06)	0.0
OTHER	385	39422(4%)	3.1(1.5) 1.96(0.24)	18.0(5.1) 2.07(0.05)	19.5(5.1) 2.30(0.08)	14.5(4.0) 2.23(0.09)	12.7(4.4) 2.20(0.07)	14.9(2.0) 2.09(0.08)	17.4(2.8) 2.12(0.06)	0.0
PARENTAL EDUCATION										
NOT GRADUATED H.S.	1267	159736(5%)	8.7(2.5) 1.94(0.07)	18.4(4.6) 1.87(0.03)	4.3(1.1) 2.00(0.12)	7.8(2.1) 2.01(0.05)	6.0(2.0) 2.02(0.07)	15.6(2.8) 1.99(0.06)	39.0(2.6) 2.03(0.03)	0.0
GRADUATED H.S.	3675	479173(3%)	6.9(1.4) 2.14(0.04)	10.6(2.5) 2.01(0.02)	8.6(1.9) 2.19(0.04)	8.4(2.1) 2.14(0.04)	10.1(2.9) 2.14(0.03)	14.9(1.6) 2.17(0.02)	40.4(2.2) 2.17(0.02)	0.0
POST H.S.	5312	740485(3%)	3.4(0.7) 2.24(0.05)	7.2(1.6) 2.09(0.03)	24.6(4.1) 2.32(0.02)	9.2(2.4) 2.24(0.04)	10.7(2.9) 2.25(0.03)	18.8(1.7) 2.27(0.02)	26.0(1.8) 2.28(0.02)	0.0
UNKNOWN	290	37087(7%)	7.0(2.2) 1.92(0.12)	23.9(6.4) 1.93(0.06)	7.3(2.4) 2.03(0.11)	11.8(3.3) 2.02(0.13)	12.7(4.0) 2.08(0.09)	13.0(2.9) 1.97(0.14)	24.2(3.0) 2.00(0.06)	0.0
AGE										
16 OR YOUNGER	1102	184232(6%)	4.7(1.3) 2.13(0.06)	10.5(3.3) 2.03(0.04)	22.0(4.3) 2.30(0.03)	11.0(3.2) 2.20(0.06)	10.5(3.0) 2.21(0.04)	17.6(2.9) 2.25(0.05)	23.8(2.8) 2.28(0.04)	0.0
17 YEARS OLD	7919	963071(1%)	5.3(1.2) 2.16(0.04)	7.8(1.8) 2.04(0.03)	17.2(2.8) 2.30(0.03)	8.4(2.1) 2.21(0.03)	10.8(3.0) 2.20(0.02)	17.2(1.3) 2.22(0.02)	33.2(1.5) 2.21(0.01)	0.0
18 OR OLDER	1636	282938(3%)	6.8(1.7) 2.04(0.05)	16.8(3.5) 1.94(0.03)	11.8(2.2) 2.19(0.06)	8.9(2.4) 2.06(0.03)	6.7(2.0) 2.11(0.06)	15.2(2.4) 2.14(0.05)	33.8(2.5) 2.10(0.03)	0.0

Table 15(67)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PARENTAL EDUCATION

	N	WEIGHTED N	NOT HS	GRAD HS	POST HS	UNKNOWN	%-OMIT
-- TOTAL --	10544	1416480(1%)	11.3(0.6) 1.99(0.02)	33.8(1.2) 2.15(0.01)	52.3(1.3) 2.27(0.01)	2.6(0.2) 1.99(0.03)	1.0
SEX							
MALE	5150	706374(2%)	10.5(0.5) 1.90(0.03)	33.9(1.2) 2.05(0.01)	52.5(1.4) 2.16(0.01)	3.1(0.2) 1.93(0.04)	1.1
FEMALE	5394	710106(2%)	12.0(0.8) 2.06(0.03)	33.7(1.4) 2.25(0.02)	52.1(1.5) 2.37(0.01)	2.2(0.3) 2.07(0.06)	0.8
ETHNICITY/RACE							
WHITE	7789	1065566(1%)	8.1(0.6) 2.07(0.03)	34.6(1.3) 2.19(0.01)	55.7(1.6) 2.30(0.01)	1.6(0.2) 2.05(0.05)	1.1
BLACK	1472	204531(2%)	15.7(1.5) 1.85(0.04)	36.4(2.1) 1.98(0.03)	43.2(2.3) 2.08(0.02)	4.6(0.6) 1.94(0.06)	0.6
HISPANIC	900	107100(3%)	35.8(4.6) 1.91(0.04)	24.7(2.7) 2.02(0.03)	31.7(3.3) 2.10(0.03)	7.8(1.4) 1.89(0.07)	0.1
OTHER	383	39283(4%)	8.1(1.2) 2.00(0.11)	23.8(2.6) 2.06(0.07)	61.2(3.5) 2.23(0.04)	6.9(1.5) 2.06(0.12)	0.4
PARENTAL EDUCATION							
NOT GRADUATED H.S.	1267	159736(5%)	100.0(0.0) 1.99(0.02)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
GRADUATED H.S.	3675	479173(3%)	0.0(0.0) *****(0.0)	100.0(0.0) 2.15(0.01)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0
POST H.S.	5312	740485(3%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 2.27(0.01)	0.0(0.0) *****(0.0)	0.0
UNKNOWN	290	37087(7%)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	0.0(0.0) *****(0.0)	100.0(0.0) 1.99(0.03)	0.0
AGE							
16 OR YOUNGER	1092	182874(6%)	6.9(0.9) 2.07(0.08)	30.3(2.0) 2.16(0.03)	60.8(2.3) 2.29(0.02)	2.0(0.4) 2.07(0.15)	0.7
17 YEARS OLD	7834	953727(1%)	9.3(0.6) 2.02(0.02)	34.4(1.3) 2.17(0.01)	54.2(1.4) 2.28(0.01)	2.1(0.2) 1.99(0.04)	1.0
18 OR OLDER	1618	279879(3%)	20.8(1.5) 1.92(0.03)	34.2(1.3) 2.08(0.02)	40.2(1.7) 2.18(0.03)	4.7(0.6) 1.96(0.08)	1.1

Table 15(68)

NAEP 1983-84 READING AND WRITING ASSESSMENT - STUDENT QUESTIONNAIRE - 11TH GRADERS
WEIGHTED RESPONSE PERCENTAGES AND GENERAL WRITING A.R.M. MEANS - REPORTING VARIABLES

PERCENT AT OR ABOVE ANCHOR POINTS

	N	WEIGHTED N	1.0	2.0	3.0	4.0
-- TOTAL --	10657	1430241(1%)	99.6(0.1)	66.3(0.8)	3.4(0.2)	0.0(0.0)
SEX						
MALE	5215	714418(2%)	99.3(0.1)	57.6(1.0)	1.8(0.2)	0.0(0.0)
FEMALE	5442	715823(2%)	99.8(0.0)	75.0(0.9)	4.9(0.4)	0.0(0.0)
ETHNICITY/RACE						
WHITE	7892	1077899(1%)	99.7(0.1)	71.1(0.9)	4.0(0.3)	0.0(0.0)
BLACK	1478	205670(2%)	99.3(0.3)	48.7(1.2)	1.2(0.3)	0.0(0.0)
HISPANIC	902	107250(3%)	98.8(0.3)	51.8(2.0)	1.0(0.3)	0.0(0.0)
OTHER	385	39422(4%)	99.8(0.2)	67.2(2.2)	3.8(1.2)	0.0(0.0)
PARENTAL EDUCATION						
NOT GRADUATED H.S.	1267	159736(5%)	99.0(0.3)	48.6(2.0)	0.9(0.3)	0.0(0.0)
GRADUATED H.S.	3675	479173(3%)	99.5(0.1)	62.8(1.0)	2.5(0.3)	0.0(0.0)
POST H.S.	5312	740485(3%)	99.8(0.1)	73.4(1.0)	4.6(0.4)	0.0(0.0)
UNKNOWN	290	37087(7%)	99.3(0.7)	46.4(3.4)	0.5(0.3)	0.0(0.0)
AGE						
16 OR YOUNGER	1102	184232(6%)	99.7(0.1)	70.7(1.8)	4.0(0.7)	0.0(0.0)
17 YEARS OLD	7919	963071(1%)	99.6(0.1)	68.5(0.9)	3.6(0.3)	0.0(0.0)
18 OR OLDER	1636	282938(3%)	99.6(0.2)	56.1(1.9)	2.0(0.4)	0.0(0.0)

APPENDIX A

Assessment Items

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
1.	N001101	PICTURE:CEREAL WITH TOY INSIDE IS PAX	H-05	3-07	H-06			
2.	N001201	LONG DIST:RATE ON CALL-LOWER EVENING RATE			H-07	3-26		3-26
3.	N001202	LONG DIST:PERSON CALLS DIFF-OPR ASSISTED			H-08	3-27		3-27
4.	N001301	KOLA COUPON:GOOD FOR ANY SIZE CARTON			H-09		H-10	
5.	N001302	KOLA COUPON:USE ON NOV. 10, 1970			H-10		H-11	
6.	N001303	KOLA COUPON:PAYMENT IS 12 CENTS			H-11		H-12	
7.	N001401	VERSE:DECK OF CARDS DESCRIBED IN POEM			H-12	3-21	H-13	3-21
8.	N001501	NUTS: DEVIL PUT PEARL IN WALNUT	H-10	3-01	H-13		H-14	
9.	N001502	NUTS: FARM WIFE WAS CLEVER AND PRACTICAL	H-11	3-02	H-14		H-15	
10.	N001503	NUTS: WANTED TRICK SOMEONE INTO CRACKING WALNUTS	H-12	3-03	H-15		H-16	
11.	N001504	NUTS: PLAN WRONG-WOMAN WAS TOO CLEVER FOR HIM	H-13	3-04	H-16		H-17	
12.	N001505	NUTS: IS THIS A GOOD STORY?	H-14	3-05	H-17		H-18	
13.	N001506	NUTS: WHY WAS THIS A GOOD OR BAD STORY	H-15	3-06	H-18		H-19	
14.	N001601	1ST AM:BITTER WINTER-EXTREMELY COLD	J-12	3-08	J-11	3-08		3-08
15.	N001602	1ST AM:ICE AGE PEOPLE DEPENDED ON ANIMALS TO LIVE	J-13	3-09	J-12	3-09		3-09
16.	N001603	1ST AM:NO LAND BRIDGE NOW-COVERED WITH WATER	J-14	3-10	J-13	3-10		3-10
17.	N001604	1ST AM:MAIN PURPOSE-EXPLN ICE AGE SETTLERS-N. AM.	J-15	3-11	J-14	3-11		3-11
18.	N001605	1ST AM:HOW INTERESTING WAS THIS ARTICLE	J-16	3-12	J-15	3-12		3-12
19.	N001606	1ST AM:HOW HARD WAS THIS ARTICLE TO READ	J-17	3-13	J-16	3-13		3-13
20.	N001701	BOOK CLUB:SHIPPING COSTS HIGHER IN CANADA			J-17	2-03	J-12	2-03
21.	N001702	BOOK CLUB:SEND NO MONEY TILL BILLED			J-18	2-04	J-13	2-04
22.	N001703	BOOK CLUB:BUY 6 MORE			J-19	2-05	J-14	2-05
23.	N001801	FLY:WANT OF THOUGHT-LACK OF THINKING	J-19	4-14	J-20	4-20		4-20
24.	N001802	FLY:FACING PROBLEMS SIMILAR TO HIS OWN	J-20	4-15	J-21	4-21		4-21
25.	N001901	CHARLEY1: MANS FEARS			J-22		J-15	
26.	N001902	CHARLEY1: MOOD OF STORY			J-23		J-16	
27.	N001903	CHARLEY1: CREATED MOOD			J-24		J-17	
28.	N002001	WISH COULD FLY:GOSSAMER CONDOR 1ST MUSCLE-POWERED	K-09	2-10	K-09	2-11	K-09	2-11
29.	N002002	WISH COULD FLY:BIKE RACER, BRYAN ALLEN FLEW CONDOR	K-10	2-11	K-10	2-12	K-10	2-12
30.	N002003	WISH COULD FLY:MACCREADY PLANE DIFF-SIMPLR/LIGHTR	K-11	2-12	K-11	2-13	K-11	2-13
31.	N002101	VIRUSES:DIFFICULT TO STUDY	K-18	4-05	K-12	4-05	K-12	4-05
32.	N002102	VIRUSES:CLOTHE IDEA-GIVE PROOF TO SUPPORT	K-19	4-06	K-13	4-06	K-13	4-06
33.	N002201	PHONE BILL:FEB 14 CALL FROM ATHENS, GA			K-14	2-14	K-14	2-14
34.	N002202	PHONE BILL:FEB 14 CALL TO ST PAUL, MN			K-15	2-15	K-15	2-15
35.	N002203	PHONE BILL:FEB 14 CALL COST $.75			K-16	2-16	K-16	2-16
36.	N002301	THE DOOR:THOUGHTS ON POEM			K-17		K-17	
37.	N002401	MOSQUITO:SIZE MOSQUITOES EXAGGERATED	L-22	2-07	L-22			
38.	N002501	MARY:WILL GET MONEY FROM NEITHER		2-13	L-23	2-17	L-27	2-17
39.	N002701	ATMOSPHERE:4 WORDS CUE-FIRST,NEXT,ABOVE,FINALLY			L-24		L-28	
40.	N002702	ATMOSPHERE:SCIENTISTS KNOW MOST ABOUT TROPOSPHERE	L-20				L-29	
41.	N002801	BETHUNE: ROOSEVELT HONOR HER BY MAKING HER DIRECTO	L-24		L-25		L-30	
42.	N002802	BETHUNE: START HER SCHOOL TO EDUCATE BLACK CHILDRN	L-25		L-26		L-31	
43.	N002803	BETHUNE: MOST IMPORTANT THINGS ABOUT HER AND WHY	L-26		L-27		L-32	
44.	N002901	SOCCER:DID YOU LIKE READING THIS ARTICLE			M-05		M-05	
45.	N002902	SOCCER:MOST POPULAR BECAUSE PLAYED BY MILLIONS			M-06		M-06	

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
46.	N002903	SOCCER:KING ED WANTED TO OUTLAW-PRACTICE ARCHERY			M-07		M-07	
47.	N002904	SOCCER:CALLED FOREIGN-IMMIGRANTS PLAYED IT MOST			M-08		M-08	
48.	N002905	SOCCER:INTRO TO ENGLISH BY ROMANS			M-09		M-09	
49.	N002906	SOCCER:PELE MASTER-FOOLED OPPONENTS BY FAKE MOVES			M-10		M-10	
50.	N003001	SUPR COURT:CONSTITUTION DESCRIPTION-BRIEF	M-10	3-14	M-11	3-15	M-11	3-15
51.	N003002	SUPR COURT:DIFFICULT RESPON FOR COURT MEMBERS	M-11	3-15	M-12	3-16	M-12	3-16
52.	N003003	SUPR COURT:"THEIR" REFERS TO PROVISIONS	M-12	3-16	M-13	3-17	M-13	3-17
53.	N003101	GOODS: DIFF TO MARKET-ROADS POOR	M-14		M-14		M-14	
54.	N003102	GOODS: YANKEE PEDDLER-TODAY SALESPERSON	M-15		M-15		M-15	
55.	N003103	GOODS: COMPARE TRADING AND SELLING IN 1700 VS. NOW	M-16		M-16		M-16	
56.	N003201	SUMMER JOB:SOC SECURITY APPLIC AT BANK OR POST OFC			N-12	1-07	N-21	1-07
57.	N003202	SUMMER JOB:BEST TIME TO FIND JOB-BEFORE MID-APRIL			N-13	1-08	N-22	1-08
58.	N003203	SUMMER JOB:NEED SS CARD TO GET HIRED			N-14	1-09	N-23	1-09
59.	N003204	SUMMER JOB:REFERENCES-PEOPLE WHO KNOW APPLICANT			N-15	1-10	N-24	1-10
60.	N003301	BOBBY:SAYS TALL IS SMART			N-16	3-19	N-25	3-19
61.	N003401	YOUNG GARDENERS:IN CENTRAL PARK-BEST			N-17			
62.	N003501	TOASTER:DRAGON/TOASTER QUALITIES COMPARED			N-18	2-10	N-27	2-10
63.	N003601	MAGIC TRICK:FIRST TIE BLACK THREAD			N-19	1-13	N-28	1-13
64.	N003602	MAGIC TRICK:DIMLY LIT RM, SAY PRODUCE FROM AIR			N-20	1-14	N-29	1-14
65.	N003701	WEB LIFE: THREAD BREAKS-FALLS APART	N-23		N-21		N-30	
66.	N003702	WEB LIFE: MAIN IDEA-PLNTS&ANMS NEED EACH OTHER	N-24		N-22		N-31	
67.	N003703	WEB LIFE: WHY YOU CHOSE A PARTICULAR MAIN IDEA	N-25		N-23		N-32	
68.	N003801	SCOTT:BEST TITLE-SCOTT'S PLAN	O-12	4-02	O-12	4-02	O-12	4-02
69.	N003802	SCOTT:6 WEEKS BETWEEN DEPOTS	O-13	4-03	O-13	4-03	O-13	4-03
70.	N003803	SCOTT:CACHE-PLACE FOR STORING THINGS	O-14	4-04	O-14	4-04	O-14	4-04
71.	N003901	SELFISH PERSON:DESCRIPTION IN PASSAGE			O-16	3-14		3-14
72.	N004001	TRIANGLE:FIGURE DRAWN			O-15			
73.	N004002	TRIANGLE:NAME FIGURE AS TRIANGLE			O-15			
74.	N004101	NONSENSE WORD 1:KAG-FIRE	O-17	4-16	O-17	4-22		4-22
75.	N004201	MEOW-WOW:2 MONTH KITTEN-FEED 3 OR 4 TMS DAILY	O-18	4-12	O-18	4-18	O-21	4-18
76.	N004202	MEOW-WOW:CAT LEAVES FOOD-LEAVE BOWL FOR HIM	O-19	4-13	O-19	4-19	O-22	4-19
77.	N004301	JAVELIN:MAIN REASON			O-20		O-23	
78.	N004302	JAVELIN:EXPLANATION OF AUTHORS IMPRESSSION			O-21		O-24	
79.	N004401	NAOMI JAMES:HOW LONG ON SAILING TRIP- 272 DAYS	P-07	4-09	P-07	4-15		4-15
80.	N004402	NAOMI JAMES:IMPORTANCE OF TRIP-BROKE WORLD RECORD	P-08	4-10	P-08	4-16		4-16
81.	N004403	NAOMI JAMES:WORST PART OF TRIP- BAD STORM	P-09	4-11	P-09	4-17		4-17
82.	N004501	AREA CODES:INFO NY-1-212-555-1212			P-10		P-20	
83.	N004502	AREA CODES:SYRACUSE 1-315-255-6011			P-11		P-21	
84.	N004601	JOBS 1900: MARTHA THINK-JOB TIRESOME			P-12		P-22	
85.	N004602	JOBS 1900: JOE FOUND HARD-STAYING IN WOODS			P-13		P-23	
86.	N004603	JOBS 1900: JOB AT HOME-ADDIE			P-14		P-24	
87.	N004604	JOBS 1900: HOW WERE THE LIVES OF THE 4 DIFFERENT			P-15		P-25	
88.	N004701	CARRIER AD:IF INTEREST & MEET REQRMNTS-CALL CIRC	Q-10	1-15	Q-07	1-15		1-15
89.	N004702	CARRIER AD:8 YR OLDS TOO YOUNG FOR JOB	Q-11	1-16	Q-08	1-16		1-16
90.	N004703	CARRIER AD:MUST DELIVER PAPERS BY 7 EACH AM	Q-12	1-17	Q-09	1-17		1-17

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
91.	N004801	SILKY 3:WISHED HE HAD SOME HAIR	Q-13	3-19	Q-10	3-20		3-20
92.	N004901	COLORADO:GOLD DISCOVERY DOESN'T BELONG	Q-14	3-17	Q-11	3-18	Q-10	3-18
93.	N005001	ARTS:BEFORE 1940 ARTS WERE ORIENTED TO ELITE			Q-13	2-06	Q-07	2-06
94.	N005002	ARTS:PRIVILEGE OF ARISTOCRATIC FEW-GREAT WORKS			Q-14	2-07	Q-08	2-07
95.	N005003	ARTS:MASS PROD NO HARM TO GENUINE ART			Q-15	2-08	Q-09	2-08
96.	N005101	DRAWING:WINNIE SHORTER THAN PAMELA-BEST STATEMENT	Q-15	2-02	Q-12	2-02		2-02
97.	N005201	TRAFFIC:APPEAR IN COURT TO PLEAD NOT GUILTY			Q-16	4-23	Q-11	4-23
98.	N005202	TRAFFIC:FINE-$3.00			Q-17	4-24	Q-12	4-24
99.	N005203	TRAFFIC:PAY FINE BY THURS, JUNE 11			Q-18	4-25	Q-13	4-25
100.	N005301	SEALS: GET FOOD ON SHORE-FROM THEIR FAT			Q-19			
101.	N005302	SEALS: SURPRISE IN MEXICO-THOUGHT SEALS EXTINCT			Q-20			
102.	N005303	SEALS: MAIN PURPOSE-DESCRIBE SEALS			Q-21			
103.	N005304	SEALS: COME SHORE YEARLY-BIRTH TO YOUNG			Q-22			
104.	N005305	SEALS: BLUBBER MEANING-FAT			Q-23			
105.	N005401	HERO: INTERESTING ARTICLE			R-05			
106.	N005402	HERO: EASE OF READING ARTICLE			R-06			
107.	N005403	HERO: MAIN IDEA-SIMON WAS A GREAT HERO			R-07			
108.	N005404	HERO: FROM WHAT COUNTRY-VENEZUELA			R-08			
109.	N005405	HERO: TRUE-COLOMBIA ONCE SPANISH			R-09			
110.	N005406	HERO: MONEY CALLED 'BOLIVARS'			R-10			
111.	N005407	HERO: GOAL NEVER REACHED-COUNTRIES JOIN TOGETHER			R-11			
112.	N005501	BUSINESS: INTERESTING			R-12		R-12	
113.	N005502	BUSINESS: EASE OF READING			R-13		R-13	
114.	N005503	BUSINESS: MAIN PURPOSE-BUSINESS TERMS MEAN			R-14		R-14	
115.	N005504	BUSINESS: OWE 50 DOLLARS FOR BIKE-A LIABILITY			R-15		R-15	
116.	N005505	BUSINESS: EXTRA MONEY IS PROFIT			R-16		R-16	
117.	N005601	TREES: TRAPS POLLUTANTS-LEAVES			R-17			
118.	N005602	TREES: CLEANING THE AIR-FILTERING PARTICLES			R-18			
119.	N005603	TREES: PURPOS GREEN BELT-REDUCE CITY POLLUTION			R-19			
120.	N005701	GRAPH:MOST POWER 1980,1985,2000-PETROLEUM			S-19	3-28	S-19	3-28
121.	N005702	GRAPH:IN 2000,HYDROPOWER SUPPLY LESS THAN COAL			S-20	3-29	S-20	3-29
122.	N005703	GRAPH:IN 2000 NUCLEAR POWER MORE % TOTAL THAN 1971			S-21	3-30	S-21	3-30
123.	N005801	ENGLISH DIC:BOOK TELLS WORD MEANINGS-DICTIONARY	S-19	1-18	S-22	1-34	S-22	1-34
124.	N005901	CARDCAT:CALL NUMBER-WRITE-IN 629.1 OB2			S-23		S-23	
125.	N005902	CARDCAT:PICTURES INDIC BY "ILLUS"			S-24		S-24	
126.	N006001	PHONE DIR:STORES SELL MILK LISTED UNDER DAIRIES			S-25		S-25	
127.	N006002	PHONE DIR:HENDRICKS MINING ON 63RD ST, 443-1502			S-26		S-26	
128.	N006003	PHONE DIR:STAR TRACKER OPEN TO REPAIR MICROSCOPE			S-27		S-27	
129.	N006101	WIND SYMBOLS:FOR 35 KNOTS-SYMBOL 3			S-30	1-18	S-30	1-18
130.	N006201	INDEX:FIND KING DARIUS INFO ON PG 23			S-31	1-29	S-31	1-29
131.	N006202	INDEX:FIND CUNEIFORM PRONUNCIATION			S-32	1-30	S-32	1-30
132.	N006203	INDEX:1875 FRENCH CONSTITUTION INFO ON PG 233			S-33	1-31	S-33	1-31
133.	N006204	INDEX:ALTERNATE HDG./DUTCH EAST INDIES-INDONESIA			S-34	1-32	S-34	1-32
134.	N006205	INDEX:DISARMAMENT IN EASTERN EUROPE INFO ON PG 279			S-35	1-33	S-35	1-33
135.	N006301	CLOTHES SIZES:SHOE SIZE 8-40-1			S-36	1-11	S-36	1-11

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
136.	N006302	CLOTHES SIZES:38 SWEATER-44			S-37	1-12	S-37	1-12
137.	N006401	TEXTS:BEST PLACE TO LOCATE BULL RUN HSTY-INDEX			S-28		S-28	
138.	N006402	TEXTS:BEST PLACE FIND DELTA DEFIN./GEOG-GLOSSARY			S-29		S-29	
139.	N006501	FIND GUIDE:OPTIONAL BETWEEN OPPRESS-ORACLE	T-26	1-19	T-26	1-19	T-26	1-19
140.	N006601	TABLE CONTENTS:MOST USEFUL IN AMERICAN HIST COURSE			T-19	1-20	T-19	1-20
141.	N006602	TABLE CONTENTS:AMERICAN INDEPENDENCE IN UNIT I			T-20	1-21	T-20	1-21
142.	N006603	TABLE CONTENTS:RECONSTRUCTION AFT CIVIL WAR-CHAP.6			T-21	1-22	T-21	1-22
143.	N006604	TABLE CONTENTS:MAJOR TOPIC CHAP.17-HAPPENINGS WWII			T-22	1-23	T-22	1-23
144.	N006605	TABLE CONTENTS:MIDDLE EAST MAP,1958-1970 ON PG.594			T-23	1-24	T-23	1-24
145.	N006701	SCIENCE INDEX:WOLVES FIRST IN BOOK			T-27	4-26	T-27	4-26
146.	N006801	MAP:SPANISH IN SOUTH			T-24		T-24	
147.	N006802	MAP:PEOPLE SETTLED IN ALASKA-NOT ENOUGH INFORM			T-25		T-25	
148.	N006901	NEWS:TV SCHEDULE-PG 22			T-36		T-36	
149.	N006902	NEWS:WEATHER FORECAST-PG 12			T-37		T-37	
150.	N006903	NEWS:STOCK AVERAGES-PGS 29-31			T-38		T-38	
151.	N007001	CATALOG CD:WHAT INFO GIVES LOCATION-GV 885 C624			T-28	1-25	T-28	1-25
152.	N007002	CATALOG CD:PG FOR OTHER BOOKS SAME TOPIC-221			T-29	1-26	T-29	1-26
153.	N007003	CATALOG CD:AUTHORS OF BOOK-COOPER & SIEDENTOP			T-30	1-27	T-30	1-27
154.	N007004	CATALOG CD:OTHER HEADING TO FIND BOOK-SIEDENTOP			T-31	1-28	T-31	1-28
155.	N007101	BUS SCHED:LAST BUS IN EVENING LEAVE CITADEL 6:45PM			T-32	3-22	T-32	3-22
156.	N007102	BUS SCHED:2ND SAT AM BUS ARRIVE DOWNTOWN 8:15AM			T-33	3-23	T-33	3-23
157.	N007103	BUS SCHED:MISS 2:35PM FROM HANCOCK WAIT TILL 3:35			T-34	3-24	T-34	3-24
158.	N007104	BUS SCHED:LV RUSTIC WED 9:42AM ARRV DWNTWN 10:15AM			T-35	3-25	T-35	3-25
159.	N007301	BRIDGER:KIND OF PEOPLE WERE MTN MEN-FUR TRAPPERS			U-19	1-01	U-19	1-01
160.	N007302	BRIDGER:BEST DESCRIBE STORIES-STRETCHED THE TRUTH			U-20	1-02	U-20	1-02
161.	N007303	BRIDGER:SIMILE-PONDS OF MUD BOILING LIKE MUSH			U-21	1-03	U-21	1-03
162.	N007304	BRIDGER:WHO DISCOVERED LAND NOW YELLOWSTONE-COLTER			U-22	1-04	U-22	1-04
163.	N007305	BRIDGER:SHOT MISSED ELK BECAUSE ELK OUT OF RANGE			U-23	1-05	U-23	1-05
164.	N007306	BRIDGER:HYPERBOLE- LAKES THAT HAD NO BOTTOM			U-24	1-06	U-24	1-06
165.	N007401	MEMORY: MAIN REASON WRITE-DESCRIBE DETAILS OF SUMM			U-25		U-25	
166.	N007402	MEMORY: FRONT PORCH IN SUMMER-COMFORTABLE			U-26		U-26	
167.	N007403	MEMORY: FAMILY LIFE WORD-CLOSE-KNIT			U-27		U-27	
168.	N007404	MEMORY: SYRUPY MEANING-HUMID			U-28		U-28	
169.	N007405	MEMORY: SET UP SWING-SHIPS SAIL ON OCEAN			U-29		U-29	
170.	N007406	MEMORY: DESCRIBE MOOD			U-30		U-30	
171.	N007407	MEMORY: CREATED MOOD			U-31		U-31	
172.	N007501	TRAVELS:MAN AFRAID-FEARFUL THOUGHTS,NO DANGER			V-29		V-38	
173.	N007502	TRAVELS:MOOD OF ARTICLE			V-30		V-39	
174.	N007503	TRAVELS:DESCRIBE MOOD OF ARTICLE			V-31		V-40	
175.	N007801	BASKETMAKER:WHY PEOPLE I BECOME SEDENTARY-GREW FD			W-38		W-40	
176.	N007802	BASKETMAKER:ABLE FIND REMAINS PEOPLE II-DRY CAVES			W-39		W-41	
177.	N007803	BASKETMAKER:TRUE-PEOPLE III LIVE IN LARGER COMMUN			W-40		W-42	
178.	N007804	BASKETMAKER:PEOPLE III USED PITHOUSES FOR CEREMONY			W-41		W-43	
179.	N008101	CLOSING:PUN-DOORMAN AT PLAZA HOTEL? NO			X-17	4-07	X-17	4-07
180.	N008102	CLOSING:PUN-FOR MORE THAN 50 YEARS? NO			X-18	4-08	X-18	4-08

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
181.	N008103	CLOSING:PUN-END SWINGING CAREER? YES			X-19	4-09	X-19	4-09
182.	N008104	CLOSING:PUN-JOB HAS HELPED HIM? NO			X-20	4-10	X-20	4-10
183.	N008105	CLOSING:PUN-UNLOCK SOME SECRETS? YES			X-21	4-11	X-21	4-11
184.	N008106	CLOSING:PUN-A LOT HINGES ON KINDNESS? YES			X-22	4-12	X-22	4-12
185.	N008107	CLOSING:MAIN PURPOSE-REPT SWEENEY LEAVES JOB			X-23	4-13	X-23	4-13
186.	N008108	CLOSING:TONE OF CAPTION IS CLEVER AND WITTY			X-24	4-14	X-24	4-14
187.	N008201	COW-TAIL: OGALOUSSA WAS KILLED WHILE HUNTING			Y-04	3-01	Y-06	3-01
188.	N008202	COW-TAIL: THEME-PERSON NOT DEAD TILL FORGOTTEN			Y-05	3-02	Y-07	3-02
189.	N008203	COW-TAIL: OGALOUSSA IS WISE,FAIR FATHER			Y-06	3-03	Y-08	3-03
190.	N008204	COW-TAIL: OGALOUSSA SHAVED HEAD-RETURNED FROM DEAD			Y-07	3-04	Y-09	3-04
191.	N008205	COW-TAIL: PULI GOT SWITCH-ASKED ABT FATHER MISSING			Y-08	3-05	Y-10	3-05
192.	N008206	COW-TAIL: IS THIS A GOOD STORY?			Y-09	3-06	Y-11	3-06
193.	N008207	COW-TAIL:WHY GOOD STORY			Y-10	3-07	Y-12	3-07
194.	N008601	CRICKETS: MAKE SOUNDS BY RUBBING WINGS	H-06	2-15				
195.	N008602	CRICKETS: WHICH MAKE CHIRPING SOUNDS-ONLY MALES	H-07	2-16				
196.	N008603	CRICKETS: WHERE ARE EARS - IN FRONT LEGS	H-08	2-17				
197.	N008701	PICTURE:DOG LYING ON TOP DOGHOUSE-BEST DESCRIPTION	H-09					
198.	N008801	YVONNE'S DOLL:COULDN'T FIND-UNDER PORCH	J-18					
199.	N008901	DOG:WHY DOESNT WANT-THINKS DOGS ARE PESTS	J-21					
200.	N008902	DOG:CHILD BRINGING HOME SNAKE	J-22					
201.	N008903	DOG:IS THIS A GOOD POEM	J-23					
202.	N008904	DOG:WHY IS THIS A GOOD POEM	J-24					
203.	N009001	FOLKS:WHO ARE THEY-HUMANS WHO LIVE NEARBY	K-12					
204.	N009002	FOLKS:GRAY FOX THINK-FOLKS WERE SENSIBLE	K-13					
205.	N009003	FOLKS: MAN WAS SITTING ON BENCH IN GARDEN	K-14					
206.	N009004	FOLKS:DO WHEN FOX CAME NEAR-MAN WAS POLITE	K-15					
207.	N009101	NONSENSE WORD 3:NABBIES-DOGS	K-16	3-18				
208.	N009201	PUZZLE 1:BIRD DESCRIBED IN PUZZLE	K-17	3-28				
209.	N009401	DUAL:WORD BAT-2 MEANINGS FOOLED NELL	L-23					
210.	N009601	TIMOTHY 1:SITTING ON STEPS	L-21	1-08				
211.	N009701	BOXBALL: MASSACHUSETTS TEACHER INVENTED BASKETBALL	M-05					
212.	N009702	BOXBALL:PURPOSE OF ARTICLE-HOW BASKETBALL INVENTED	M-06					
213.	N009703	BOXBALL:TRUE-FOOTBALL INVENTED BEFORE BASKETBALL	M-07					
214.	N009704	BOXBALL:AT FIRST USED PEACH BASKET FOR GOALS	M-08					
215.	N009705	BOXBALL:BOTTOMS CUT OUT-TO MAKE IT EASIER	M-09					
216.	N009801	PUZZLE 3:CHAIR DESCRIBED IN PUZZLE	N-12					
217.	N009901	DESCRIPTION 3:PERSON HAS SEEN TOY MANY TIMES	N-13					
218.	N010001	DOG & SHADOW:LIKED READING IT	N-17	1-05				
219.	N010002	DOG & SHADOW: SAW HIMSELF IN THE STREAM	N-18	1-06				
220.	N010003	DOG & SHADOW: TEACHES LESSON-GREED DOESN'T PAY	N-19	1-07				
221.	N010101	SANDWICH:LIKED READING IT	N-20	1-09				
222.	N010102	SANDWICH:NAMED AFTER PERSON WHO INVENTED IT	N-21	1-10				
223.	N010103	SANDWICH:WANTED MEAT IN BREAD TO EAT AND GAMBLE	N-22	1-11				
224.	N010201	DESCRIPTION 1:CLOWN DESCRIBED IN PASSAGE	O-16	3-20				
225.	N010301	SNOWMAN:BEST DESCRIPTION-SOMEONE MADE SNOWMAN	O-15	2-09				

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
226.	N010401	TOOTH TROUBLE: SPEAKER-CHILD	O-20					
227.	N010402	TOOTH TROUBLE: TRUE-PULLED LOOSE TOOTH	O-21					
228.	N010403	TOOTH TROUBLE: "NOT ME, WONT PRDCE IT" SAME-NO GRO	O-22					
229.	N010501	QUICKSAND:HOW TEST FOR IT-POKE WITH A STICK	P-10	2-03				
230.	N010502	QUICKSAND:MAIN PURPOSE-TO TELL WAYS AVOID DANGER	P-11	2-04				
231.	N010503	QUICKSAND:IT IS SOUPY SAND YOU CAN'T STAND ON	P-12	2-05				
232.	N010504	QUICKSAND:IF STEP IN,LIE ON BACK & STRETCH OUT ARM	P-13	2-06				
233.	N010601	THAD:CANDIDATES FOR PRES NOT ALLOWED GIVE GIFTS	P-14					
234.	N010602	THAD:MAGGIE THOUGHT THAD GOOD BUT NEED HER HELP	P-15					
235.	N010603	THAD:MASSIVE STAMPEDE-LOT OF PEOPLE RUSHING	P-16					
236.	N010604	THAD:EXAGGERATED-CAN DO EVERYTHING IN YELLOW PAGES	P-17					
237.	N010605	THAD:MAGGIE FIRST HELPED THAD WITH SPEECH	P-18					
238.	N010701	SENTENCE 3:MOST SENSE-BALL ROLLED DOWN THE STREET	P-19					
239.	N010801	ANGRY: CHILD COMES OUT WHEN FEELS BETTER	Q-16	3-29				
240.	N010901	STARS UNSEEN: LIKED READING IT	Q-17					
241.	N010902	STARS UNSEEN:STAR BECOMES DEAD BY USING UP FUEL	Q-18					
242.	N010903	STARS UNSEEN:MAIN IDEA-STARS EXIST-WE CAN'T SEE	Q-19					
243.	N010904	STARS UNSEEN:GRAVITY OF DEAD STARS-PUSH & PULL	Q-20					
244.	N011001	REPORTER: WHO WAS ERNIE PYLE-NEWSPAPER REPORTER	R-05					
245.	N011002	REPORTER:HOW PYLES WRITING CHANGE-TROOP MVMNT&GNRL	R-06					
246.	N011003	REPORTER:HAPPENED TO PYLE-FAMOUS REPORTER	R-07					
247.	N011004	REPORTER:WHY PYLE CHANGE NEWS-REMEMBED DEATH SOLDI	R-08					
248.	N011101	KIND OF BK:ATMOSPHERE FROM SCIENCE BOOK	R-09	1-36				
249.	N011201	DOGS'QUAL:BITTEN BY DOG, DISAGREE	R-10	2-18				
250.	N011301	SKUNK CABBAGE:NAME-SMELLS LIKE SKUNK,LOOKS CABBAGE	R-11	3-21				
251.	N011302	SKUNK CABBAGE:HARD TO SEE-HIDDEN UNDER HOOD	R-12	3-22				
252.	N011401	BREATHING:TRUE-BLOOD MOVES OXYGEN	R-13					
253.	N011402	BREATHING: HOW AIR MOVES TO LUNGS-THROUGH WINDPIPE	R-14					
254.	N011403	BREATHING:FUNCTION OF AIR SACS IN LUNGS-O2 FROM LU	R-15					
255.	N011404	BREATHING:CO2 FORM IN BODY-CELLS FORM CO2 WASTE	R-16					
256.	N011501	DICTIONARY:TO FIND WORD MEANING-DICTIONARY BEST	T-27					
257.	N011601	DICTIONARY:DEFINITION TOME-A LARGE BOOK	S-21	3-23				
258.	N011602	DICTIONARY:TOMORROW SYLLABICATED-TO MOR ROW	S-22	3-24				
259.	N011603	DICTIONARY:PLURAL IS TONSILLECTOMIES	S-23	3-25				
260.	N011604	DICTIONARY:TOLERANCE IS A NOUN	S-24	3-26				
261.	N011605	DICTIONARY:TONIC-MAKES YOU FEEL BETTER	S-25	3-27				
262.	N011701	WHICH WORD COMES FIRST IN DICTIONARY- FLEA	S-30	4-18				
263.	N011801	ENCYCLOPEDIAS 2:WASHINGTON IN VOL 11	S-31					
264.	N011901	INDEX:FIND OUT ABOUT SALMON-PGS 84 &85	S-26	1-24				
265.	N011902	INDEX:ALTERNATE INFO;RAILRDS-TRAVEL & TRANSPORT	S-27	1-25				
266.	N011903	INDEX:FIND MAP OF SNAKE RIVER-PG 84	S-28	1-26				
267.	N011904	INDEX:FIND MAP S. AMERICAN RAIN FORESTS-PG. 119	S-29	1-27				
268.	N012001	DECLARATION OF INDEPENDENCE: BEST INFO ENCYCLOPEDI	S-32					
269.	N012101	CODE:WHAT DOES HPPE ACTUALLY SPELL-GOOD	S-33					
270.	N012201	DICTIONARY:PLUME IS FEATHER	T-19	1-28				

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
271.	N012202	DICTIONARY:MORE THAN 1 PLOWMAN IS PLOWMEN	T-20	1-29				
272.	N012203	DICTIONARY:PLUNDER-ROB	T-21	1-30				
273.	N012204	DICTIONARY:PLUM-IMPORTANT WORK	T-22	1-31				
274.	N012301	MUSHROOM: 3 PARTS-CAP, STEM, GILLS	S-20	1-12				
275.	N012401	INDEX:ALPHA LIST OF TOPICS AND PAGE NUMBERS	T-24					
276.	N012501	WHALE FOOD: INFO FOUND IN ENCYCLOPEDIA	T-25					
277.	N012601	ROTOR:BEST PLACE FIND INFO-DICTIONARY ROTOR	T-23	1-13				
278.	N012701	ENCYCLOPEDIA:INFO ON MEXICO IN VOLUME 6	T-28	1-32				
279.	N012702	ENCYLOPEDIA:INFO ON INVENTIONS OF EDISON IN VOL.3	T-29	1-33				
280.	N012703	ENCYLOPEDIA:INFO ON IOWA FARM PRODUCTS IN VOL. 5	T-30	1-34				
281.	N012704	ENCYCLOPEDIA:INFO ON N.Y.RIVERS & LAKES IN VOL. 7	T-31	1-35				
282.	N012801	GRAPH:SPENT MOST ON A BOOK	T-32	1-20				
283.	N012802	GRAPH:RECORD COST $2.50	T-33	1-21				
284.	N012803	GRAPH: 5 ITEMS COST MORE THAN PAINTBRUSH	T-34	1-22				
285.	N012804	GRAPH:SPENT SAME AMOUNT ON PAINTS,BIKE PARTS	T-35	1-23				
286.	N012901	TIMOTHY:3 TEENAGERS TALKING ABOUT HEAT	U-19					
287.	N013001	OIL SPILL: WHAT IS THE GEORGIA-A SHIP	U-20					
288.	N013002	OIL SPILL: WHERE WAS SPILL-5 MILES FROM BEACH	U-21					
289.	N013003	OIL SPILL:WHY LOGS NO STOP OIL-HIGH WAVES	U-22					
290.	N013004	OIL SPILL: WHAT ARE PEOPLE ASKED TO DO-CLEAN BEACH	U-23					
291.	N013101	THE COLD:BOY LEFT SHADOW-FROZE TO SIDE OF HOUSE	U-24	1-01				
292.	N013102	THE COLD:GIRLS FIGHT WITH MELTED WORDS	U-25	1-02				
293.	N013103	THE COLD:DUCKS FLY AWAY WITH POND	U-26	1-03				
294.	N013104	THE COLD:WRITER MAKES STORY SOUND PLAYFUL & FUNNY	U-27	1-04				
295.	N013201	BULLFIGHT:BULL CHARGES CAPE MOTION	V-29	4-17				
296.	N013301	DESCRIPTION 2:UNHAPPY PERSON DESCRIBED IN PASSAGE	V-30	1-14				
297.	N013401	FROM THE PLANET:BOTCHIK FELT ANNOYED AND UPSET	V-31					
298.	N013402	FROM THE PLANET:THOUGHT NO LIFE-THICK CLOUD COVER	V-32					
299.	N013403	FROM THE PLANET:IN GLASS CAGE WAS A HUMAN BEING	V-33					
300.	N013501	CRIME:HARD TO PROVE OWN BIKE IF REPAINTED & # GONE	W-37	4-07				
301.	N013502	CRIME:MAIN PURPOSE-TO GIVE SECRET WAY TO MARK BIKE	W-38	4-08				
302.	N013601	SWINGING/STAR:PEOPLE LIKE PIG IF LAZY AND RUDE	W-40					
303.	N013602	SWINGING/STAR:PEOPLE SHOULD DIFFER-TRY BE BETTER	W-41					
304.	N013603	SWINGING/STAR: LINE 4 DOESN'T RHYME WITH OTHERS	W-42					
305.	N013701	OLD MAN:STORY TELLS HOW MAN LOOKS	X-17					
306.	N013901	SAVING ENERGY: MAIN IDEA-CONSERVE OIL & NAT GAS	X-18					
307.	N013902	SAVING ENERGY: SOURCE MOST ENERGY-OIL & NAT GAS	X-19					
308.	N013903	SAVING ENERGY: WHAT CAN SOLAR ENERGY PROVIDE-HEAT	X-20					
309.	N014001	NONSENSE WORD 3:TUP-PAPER	M-13	2-14				
310.	N014101	SENTENCE 1:MOST SENSE-BLEW HOUSE DOWN	Q-21					
311.	N014201	TIMOTHY 2:TEENAGERS STANDING IN CIRCLES	V-34					
312.	N014301	FRONTIER WOMEN: BEST DESCRIBES WOMEN-WORKED HARD	N-14					
313.	N014302	FRONTIER WOMEN: ACTIVITIES PERFORMED-MAKE TOOLS&PL	N-15					
314.	N014303	FRONTIER WOMEN: MADE FROM ANML HORNS/BONES-TOOLS	N-16					
315.	N014501	CONNECT DOTS:ALONG LINE,CONNECT DOTS	V-35					

Table A(1)
Reading Items and Locations

N	NAEP ID	DESCRIPTION	Grade 4/Age 9		Grade 8/Age 13		Grade 11/Age 17	
			Block	Tape	Block	Tape	Block	Tape
316.	N014502	CONNECT DOTS:DRAW LINE TO TOUCH CIRCLES	V-35					
317.	N014503	CONNECT DOTS:WRITE 3 IN EACH CIRCLE	V-35					
318.	N015101	BLACK ELK: THINK WASICHUS WERE GREEDY					R-17	
319.	N015102	BLACK ELK: WHO WERE THE WASICHUS					R-18	
320.	N015103	BLACK ELK:DRINKS WATERS DREAM PREDICT					R-19	
321.	N015104	BLACK ELK: MAIN PURPOSE OF STORY					R-20	
322.	N015201	PEOPLE LEARN TO READ: IN SCHOOL					N-26	
323.	N015501	CHAMONIX: LIKED READING IT					P-15	
324.	N015502	CHAMONIX: WHY SO LONG TO REACH-WINDS TOO STRONG					P-1	
325.	N015503	CHAMONIX: DEVOUASSOU-MAN WHO FOUND CLIMBERS					P-17	
326.	N015504	CHAMONIX: DESMAISON SURVIVE BY MENTAL/PHYS STRNGTH					P-18	
327.	N015505	CHAMONIX: WHY DESMAISON CRY-OVERCOME SUFFERING,JOY					P-19	
328.	N015901	HIGH TECH PIZZA: WHY PIZZA					Q-14	
329.	N015902	HIGH TECH PIZZA: INTERMEDIATE STAGE					Q-15	
330.	N015903	HIGH TECH PIZZA: CORNSTARCH USED					Q-16	
331.	N015904	HIGH TECH PIZZA: DESCRIBE FABRICATION OF PIZZA					Q-17	
332.	N016001	VOTING: MAIN PURPOSE					O-15	
333.	N016002	VOTING: MEANING OF SUFFRAGE					O-16	
334.	N016003	VOTING: FIRST CONGRESSWOMEN					O-17	
335.	N016004	VOTING: DISASTER AT TRIANGLE SHIRTWAIST					O-18	
336.	N016005	VOTING: ROSE SCHNEIDERMAN SAY					O-19	
337.	N016006	VOTING: WW I HELPED SUFFRAGIST CAUSE					O-20	
338.	N017001	THE CHIP: MAIN IDEA					H-07	
339.	N017002	THE CHIP: WIDESPREAD RESULT					H-08	
340.	N017003	THE CHIP: MEANING OF TRIFLING					H-09	

Table A(2)

Background and Attitude Items by Topic

General Background

Demographic background and home environment

Ethnicity

B000101	ETHNICITY
B000102	OTHER ETHNICITY
B000201	ARE YOU HISPANIC
B000202	OTHER SPANISH-HISPANIC

Language background

B000301	WHAT LANGUAGE DO YOU SPEAK MOST OFTEN IN HOME
B000302	OTHER LANGUAGE YOU SPEAK MOST OFTEN IN HOME
B000401	WHAT LANGUAGE DO OTHERS SPEAK MOST OFTEN IN HOME
B000402	OTHER LANGUAGE OTHERS SPEAK MOST OFTEN IN HOME
B000501	FIRST OTHER LANGUAGE YOU KNOW
B000502	SECOND OTHER LANGUAGE YOU KNOW
B000503	THIRD OTHER LANGUAGE YOU KNOW

Mother work outside home

B000801	DOES YOUR MOTHER WORK OUTSIDE YOUR HOME

Parents' education

B000601	HOW FAR IN SCHOOL DID YOUR FATHER GO
B000701	HOW FAR IN SCHOOL DID YOUR MOTHER GO

Objects in home

B000901	DOES YOUR FAMILY GET A NEWSPAPER REGULARLY
B000902	IS THERE A DICTIONARY IN YOUR HOME
B000903	IS THERE AN ENCYCLOPEDIA IN YOUR HOME
B000904	ARE THERE MORE THAN 25 BOOKS IN YOUR HOME
B000905	DOES YOUR FAMILY GET MAGAZINES REGULARLY
B000906	IS THERE A VIDEO GAME IN YOU HOME
B000907	IS THERE A COMPUTER IN YOUR HOME

Mobility

B002001	HOW MAY DIFFERENT TOWNS HAVE YOU LIVED IN
S002801	WHERE DID YOU LIVE AT AGE 9
S002802	STATE
S002803	COUNTRY
S005901	WHERE DID YOU LIVE AT AGE 13

Table A(2)
(continued)

Who is home after school

- S003201 WHAT DO YOU USUALLY DO AFTER SCHOOL
- S003202 IF YOU GO HOME AFTER SCHOOL, WHO IS USUALLY THERE
- S003203 WHAT OTHER ADULT

Family composition

- S003901 HOW MANY OLDER BROTHERS AND SISTERS
- S003902 HOW MANY YOUNGER BROTHERS AND SISTERS

Educational background and plans

School program

- B001001 DO YOU HAVE GYM ONCE PER WEEK
- B001002 DO YOU HAVE ART ONCE PER WEEK
- B001003 DO YOU HAVE MUSIC ONCE PER WEEK
- B001004 DO YOU HAVE FOREIGN LANGUAGE ONCE PER WEEK
- B001005 DO YOU HAVE COMPUTER CLASS ONCE PER WEEK
- B001006 DO YOU HAVE DRAMA CLASS ONCE PER WEEK
- B001007 DO YOU HAVE SCIENCE ONCE PER WEEK

Grades

- B001901 WHICH DESCRIBES YOUR GRADES IN SCHOOL

Preschool experience

- S002701 DID YOU GO TO KINDERGARDEN
- S002702 DID YOU GO TO DAY CARE
- S002703 DID YOU GO TO NURSERY SCHOOL
- S002704 DID YOU GO TO HEADSTART

Educational expectations

- S003401 DO YOU EXPECT TO GRADUATE FROM HIGH SCHOOL

Applied to college

- S005701 HAVE YOU APPLIED TO COLLEGE

Career goals

- S005801 WHAT ARE LONG-TERM CAREER GOALS

High school program

- Science courses taken
  - S006001 HAVE YOU TAKEN GENERAL SCIENCE
  - S006002 HAVE YOU TAKEN BIOLOGY
  - S006003 HAVE YOU TAKEN CHEMISTRY
  - S006004 HAVE YOU TAKEN PHYSICS

Table A(2)
(continued)

S006005	WHAT OTHER SCIENCE COURSES

Math courses taken

S006101	HAVE YOU TAKEN GENERAL MATH 1
S006102	HAVE YOU TAKEN GENERAL MATH 2
S006103	HAVE YOU TAKEN FIRST-YEAR ALGEBRA
S006104	HAVE YOU TAKEN SECOND-YEAR ALGEBRA
S006105	HAVE YOU TAKEN GEOMETRY
S006106	HAVE YOU TAKEN CALCULUS
S006107	WHAT OTHER MATH COURSES-1
S006108	WHAT OTHER MATH COURSES-2
S006109	WHAT OTHER MATH COURSES-3

Special courses taken

S006401	EVER HAD REMEDIAL ENGLISH
S006402	EVER HAD REMEDIAL MATHEMATICS
S006403	EVER HAD HONORS ENGLISH
S006404	EVER HAD HONORS MATH
S006405	EVER HAD HONORS SCIENCE
S006406	EVER HAD BILINGUAL PROGRAM
S006407	EVER HAD FAMILY-LIFE OR SEX EDUCATION
S006408	EVER HAD ALCOHOL OR DRUG-ABUSE EDUCATION
S006409	EVER HAD PROGRAM BECAUSE OF PHYSICAL PROBLEM
S006410	EVER HAD PROGRAM BECAUSE OF SPEECH PROBLEM
S006501	HAVE YOU TAKEN AGRICULTURE
S006502	HAVE YOU TAKEN AUTO MECHANICS
S006503	HAVE YOU TAKEN COMMERCIAL ARTS
S006504	HAVE YOU TAKEN COMPUTER PROGRAMMING
S006505	HAVE YOU TAKEN CARPENTRY
S006506	HAVE YOU TAKEN ELECTRICAL CONSTRUCTION
S006507	HAVE YOU TAKEN MASONRY
S006508	HAVE YOU TAKEN PLUMBING
S006509	HAVE YOU TAKEN COSMETOLOGY
S006510	HAVE YOU TAKEN DRAFTING
S006511	HAVE YOU TAKEN ELECTRONICS
S006512	HAVE YOU TAKEN HOME ECONOMICS
S006513	HAVE YOU TAKEN MACHINE SHOP
S006514	HAVE YOU TAKEN MEDICAL OR DENTAL ASSIST
S006515	HAVE YOU TAKEN PRACTICAL NURSE
S006516	HAVE YOU TAKEN FOOD SERVICE
S006517	HAVE YOU TAKEN SALES OR MERCHANDISING
S006518	HAVE YOU TAKEN SECRETARIAL
S006519	HAVE YOU TAKEN WELDING
S006520	WHAT OTHER COURSES HAVE YOU TAKEN

Table A(2)
(continued)

Plans after high school

S006601	WHAT ONE THING WILL YOU DO AFTER HIGH SCHOOL
S006701	WHAT OTHER THINGS WILL YOU DO AFTER HIGH SCHOOL

Computer exposure and use

Computer exposure and use

S002901	DO YOU USE A COMPUTER AT HOME
S002902	DO YOU USE A COMPUTER AT THE LIBRARY
S002903	DO YOU USE A COMPUTER AT A FRIENDS HOUSE
S002904	HOW OFTEN DO YOU USE A COMPUTER AT SCHOOL
S003001	DO YOU USE A COMPUTER TO PLAY GAMES
S003002	DO YOU USE A COMPUTER TO LEARN THINGS
S003003	DO YOU USE A COMPUTER TO WRITE STORIES OR PAPERS
S003101	HOW OFTEN DO YOU WRITE COMPUTER PROGRAMS

Use of time

Time spent on homework

B001701	HOW MUCH TIME DID YOU SPEND ON HOMEWORK YESTERDAY

TV watching

B001801	HOW MUCH TELEVISION DO YOU WATCH EACH DAY

How much free time

S006801	HOW MUCH FREE TIME ON AVERAGE SCHOOL DAY

Use of free time

S005001	WHEN FREE TIME, HOW OFTEN WATCH TV
S005002	WHEN FREE TIME, HOW OFTEN READ A BOOK
S005003	WHEN FREE TIME, HOW OFTEN WRITE IN DIARY
S005004	WHEN FREE TIME, HOW OFTEN CALL A FRIEND
S005005	WHEN FREE TIME, HOW OFTEN BE WITH FRIENDS
S005006	WHEN FREE TIME, HOW OFTEN GO SHOPPING
S005007	WHEN FREE TIME, HOW OFTEN PLAY A SPORT
S005008	WHEN FREE TIME, HOW OFTEN GO HUNTING OR FISHING
S005009	WHEN FREE TIME, HOW OFTEN TAKE A WALK
S005010	WHEN FREE TIME, HOW OFTEN WORK AT A COMPUTER
S005011	WHEN FREE TIME, HOW OFTEN PLAY VIDEO GAMES
S005012	WHEN FREE TIME, HOW OFTEN READ A NEWSPAPER
S005013	WHEN FREE TIME, HOW OFTEN GET A SNACK
S005014	WHEN FREE TIME, HOW OFTEN DO EXTRA HOMEWORK
S005015	WHEN FREE TIME, HOW OFTEN WRITE A LETTER
S005016	WHEN FREE TIME, HOW OFTEN LISTEN TO MUSIC

Table A(2)
(continued)

S005017	WHEN FREE TIME, HOW OFTEN DO SOMETHING ELSE
S005018	WHEN FREE TIME, WHAT IS IT
S005019	WHEN FREE TIME WHAT ACTIVITY SPEND MOST TIME

Activities

S003601	HOW OFTEN DO YOU GO TO A MOVIE
S003602	HOW OFTEN DO YOU GO TO A PLAY
S003603	HOW OFTEN DO YOU GO TO A CONCERT
S003604	HOW OFTEN DO YOU GO TO A PARTY
S003605	HOW OFTEN DO YOU GO TO PUBLIC LIBRARY
S003606	HOW OFTEN DO YOU TRAVEL TO A PLACE AWAY FROM HOME
S003607	HOW OFTEN DO YOU GO SHOPPING
S003608	HOW OFTEN DO YOU GO TO A SPORTS EVENT
S003609	HOW OFTEN DO YOU PLAY CARD OR TABLE GAMES
S003610	HOW OFTEN DO YOU VISIT RELATIVES
S003611	HOW OFTEN DO YOU GO TO A MUSEUM
S003612	HOW OFTEN DO YOU GO CAMPING
S003613	HOW OFTEN DO YOU STAY HOME ALONE
S003614	WHAT ACTIVITY DO YOU DO MOST OFTEN

Orientation to school

Bored

S003701	DO YOU EVER FEEL BORED AT SCHOOL

Sanctions

S003801	DURING PAST YEAR HOW OFTEN SENT TO PRINCIPALS OFF
S003802	DURING PAST YEAR HOW OFTEN PLACED ON PROBATION
S003803	DURING PAST YEAR HOW OFTEN GIVEN A DETENTION
S003804	DURING PAST YEAR HOW OFTEN WARNED ABOUT ATTENDANC
S003805	DURING PAST YEAR HOW OFTEN WARNED ABOUT GRADES
S003806	DURING PAST YEAR HOW OFTEN WARNED ABOUT BEHAVIOR

Absenteeism

S004001	HOW MANY DAYS OF SCHOOL MISSED LAST MONTH

Lateness

S004101	HOW MANY TIMES LATE FOR SCHOOL LAST MONTH

Ratings of school

S006201	RATE SCHOOL:PREPARING FOR COLLEGE
S006202	RATE SCHOOL:PREPARING FOR CAREER
S006203	RATE SCHOOL:PREPARING FOR LIFE
S006204	RATE SCHOOL:EXTRACURRICULAR ACTIVITIES-VARIETY
S006205	RATE SCHOOL:EXTRACURRICULAR ACTIVITIES-QUALITY

Table A(2)
(continued)

S006206	RATE SCHOOL:FACULTY INTEREST
S006207	RATE SCHOOL:QUALITY OF FACULTY
S006208	RATE SCHOOL:QUALITY OF STUDENT LIFE

Ratings of own school experience

S006301	TRUE OR FALSE: SATISFIED WITH PROGRESS OF EDUCAT
S006302	TRUE OR FALSE: NOT LEARNING WHAT NEED TO KNOW
S006303	TRUE OR FALSE: HAD DISCIPLINARY PROBLEMS IN PAST
S006304	TRUE OR FALSE: AM INTERESTED IN SCHOOL
S006305	TRUE OR FALSE: EVERY ONCE IN A WHILE CUT A CLASS
S006306	TRUE OR FALSE: DO NOT FEEL SAFE AT THIS SCHOOL
S006307	TRUE OR FALSE: WISH COULD GO TO DIFFERENT SCHOOL

Reading and Writing Background

Student perceptions of instructional practices in reading and writing

School writing assignments

B002401	REPORTS AND ESSAYS WRITTEN FOR SCHOOL LAST 6 WEEK
S000101	TIME SPENT IN ENGLISH CLASS LEARNING TO WRITE
S000201	REPORTS AND PAPERS WRITTEN FOR SCHOOL LAST 6 WEEK
S000301	WRITINGS DONE LAST WEEK FOR SOCIAL STUDIES CLASS
S000401	WRITINGS DONE LAST WEEK FOR SCIENCE

How teacher assists in writing

S000601	WHEN WRITING HOW OFTEN TEACHER ASKS TO MAKE NOTES
S000602	WHEN WRITING HOW OFTEN TEACHER ASKS MAKE OUTLINE
S000603	WHEN WRITING HOW OFTEN TEACHER ASKS NOTE CHANGES
S000604	WHEN WRITING HOW OFTEN TEACHER ASKS TALK DURING
S000605	WHEN WRITING HOW OFTEN TEACHER ASKS TALK MATES
S000606	WHEN WRITING HOW OFTEN TEACHER ASK REDO BEFOR GRD
S000607	WHEN WRITING HOW OFTEN TEACHER ASK REDO AFTER GRD

Teacher feedback after writing

B002604	HOW OFTEN DOES TEACHER WRITE SUGGESTIONS ON PAPER
B002605	HOW OFTEN DOES TEACHER DISCUSS FINISHED PAPERS
S001701	HOW OFTEN DOES TEACHER ASK IF YOU FOLLOWED DIRECT
S001702	HOW OFTEN DOES TEACHER ASK IF YOU WROTE ENOUGH
S001703	HOW OFTEN DOES TEACHER ASK YOUR IDEAS IN PAPER
S001704	HOW OFTEN DOES TEACHER ASK EXPLANATIONS IN PAPER
S001705	HOW OFTEN DOES TEACHER ASK EXPRESS FEELINGS PAPER
S001706	HOW OFTEN DOES TEACHER ASK ORGANIZATION IN PAPER
S001707	HOW OFTEN DOES TEACHER ASK WORDS YOU USED IN PAPE

Table A(2)
(continued)

S001708	HOW OFTEN DOES TEACHER ASK SPELLING, GRAM IN PAPE
S001709	HOW OFTEN DOES TEACHER ASK YOUR NEATNESS IN PAPER
S002501	HOW OFTEN DOES TEACHER MARK ERRORS ON PAPERS
S002502	HOW OFTEN DOES TEACHER WRITE NOTES ON PAPERS
S002503	HOW OFTEN DOES TEACHER POINT OUT GOOD THINGS
S002504	HOW OFTEN DOES TEACHER POINT OUT NOT GOOD THINGS
S002505	HOW OFTEN DOES TEACHER MAKE SUGGESTIONS FOR NEXT
S002506	HOW OFTEN DOES TEACHER SHOW INTEREST IN WRITING

Teacher behavior around reading

S004601	HOW OFTEN WITH NEW READING TEACHER POINT HARD WOR
S004602	HOW OFTEN WITH NEW READING TEACHER PREVIEW READIN
S004603	HOW OFTEN WITH NEW READING TEACHER READ PART ALOU
S004701	HOW OFTEN DOES TEACHER LIST OF QUESTS AS YOU READ
S004702	HOW OFTEN DOES TEACHER TELL HOW TO FIND MAIN IDEA
S004703	HOW OFTEN DOES TEACHER TELL HOW TO READ FASTER

Teacher behavior around writing

B002601	HOW OFTEN ENCOURAGED MAKE NOTES ON TOPIC OF PAPER
B002602	HOW OFTEN ENCOURAGED TO MAKE OUTLINES OF PAPER

Time spent learning to write

B002501	PART OF CLASS TIME SPENT LEARNING TO WRITE REPORT
B002603	HOW OFTEN DO YOU WRITE PAPER MORE THAN ONCE
B002606	HOW OFTEN DO YOU IMPROVE PAPER AFTER RETURN

Self-assessment as reader and writer

Self-assessment as reader

S003301	WHAT KIND OF READER ARE YOU

Self-assessment as writer

B002607	DO YOU ENJOY WORKING ON WRITING ASSIGNMENTS
S001201	HOW OFTEN IS TRUE: LIKE TO WRITE
S001202	HOW OFTEN IS TRUE: AM A GOOD WRITER
S001203	HOW OFTEN IS TRUE: THINK WRITING IS WASTE OF TIME
S001204	HOW OFTEN IS TRUE: PEOPLE LIKE WHAT I WRITE
S001205	HOW OFTEN IS TRUE: WRITE ON OWN AWAY FROM SCHOOL
S001206	HOW OFTEN IS TRUE: DISLIKE WRITING TO BE GRADED
S001207	HOW OFTEN IS TRUE: WOULDNT WRITE IF NOT FOR SCHOO

Table A(2)
(continued)

Student study habits and reading and writing behavior

Pages read for school and homework

B001101	HOW MANY PAGES READ IN SCHOOL AND FOR HOMEWORK

Frequency of kinds of writing

B001201	STORIES WRITTEN FOR ENGLISH LAST WEEK
B001202	ESSAYS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001203	POEMS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001204	PLAYS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001205	LETTERS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001206	BOOK REPORTS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001207	OTHER REPORTS WRITTEN FOR ENGLISH CLASS LAST WEEK
B001208	I DO NOT HAVE AN ENGLISH CLASS
S000501	WRITINGS DONE LAST WEEK NON-SCHOOL RELATED
S001901	HOW OFTEN DO YOU WRITE A BOOK REPORT
S001902	HOW OFTEN DO YOU WRITE ABOUT SCIENCE EXPERIMENT
S001903	HOW OFTEN DO YOU WRITE LETTER TO A RELATIVE
S001904	HOW OFTEN DO YOU WRITE NOTES OR MESSAGE
S001905	HOW OFTEN DO YOU WRITE STORY THAT NOT HOMEWORK

Last thing read on own

B001401	WHAT WAS THE LAST THING YOU READ FOR SCHOOL
B001501	WHAT WAS THE LAST THING YOU READ ON YOUR OWN
B001502	OTHER THING YOU READ ON YOUR OWN

Frequency of kinds of reading behavior

S004301	HOW OFTEN DO YOU READ A STORY OR NOVEL
S004302	HOW OFTEN DO YOU READ A POEM
S004303	HOW OFTEN DO YOU READ A PLAY
S004304	HOW OFTEN DO YOU READ A NEWSPAPER
S004305	HOW OFTEN DO YOU READ A MAGAZINE
S004306	HOW OFTEN DO YOU READ A SCIENCE BOOK
S004307	HOW OFTEN DO YOU READ A BIOGRAPHY
S004308	HOW OFTEN DO YOU READ A HOW-TO-DO BOOK
S004309	HOW OFTEN DO YOU READ A BOOK ABOUT OTHER TIMES
S004310	HOW OFTEN DO YOU READ A SPORTS BOOK
S004311	HOW OFTEN DO YOU READ WORDS OF SONG

Behavior around writing

S000901	WHEN WRITING HOW OFTEN ASK SELF SUBJECT PAPER
S000902	WHEN WRITING HOW OFTEN LOOK UP FACTS IN BOOKS
S000903	WHEN WRITING HOW OFTEN THINK BEFORE WRITING
S000904	WHEN WRITING HOW OFTEN THINK ABOUT LAYOUT
S000905	WHEN WRITING HOW OFTEN USE DIFF STYLES PER PERSON
S000906	WHEN WRITING HOW OFTEN MAKE CHANGES AS YOU WRITE

Table A(2)
(continued)

S000907	WHEN WRITING HOW OFTEN MAKE CHANGE AFTER WRITING
S001001	HOW OFTEN HAVE YOU SHOWN FRIENDS YOUR WRITINGS
S001002	HOW OFTEN HAVE PAPERS BEEN PRINTED IN SCHOOL PAPE
S001003	HOW OFTEN DOES YOUR FAMILY READ YOUR PAPERS
S001301	HOW OFTEN IS TRUE: MOVE SENTENCES AROUND
S001302	HOW OFTEN IS TRUE: ADD NEW IDEAS OR INFORMATION
S001303	HOW OFTEN IS TRUE: TAKE OUT UNDESIRED PARTS
S001304	HOW OFTEN IS TRUE: CHANGE WORDS
S001305	HOW OFTEN IS TRUE: CORRECT SPELLING MISTAKES
S001306	HOW OFTEN IS TRUE: CORRECT GRAMMAR MISTAKES
S001307	HOW OFTEN IS TRUE: CORRECT PUNCTUATION MISTAKES
S001308	HOW OFTEN IS TRUE: REWRITE MOST OF PAPER
S001309	HOW OFTEN IS TRUE: THROW OUT AND START OVER
S001601	HOW OFTEN DO YOU LIST THINGS TO BUY
S001602	HOW OFTEN DO YOU COPY RECIPE OR DIRECTIONS
S001603	HOW OFTEN DO YOU FILL OUT ORDER BLANKS
S001604	HOW OFTEN DO YOU KEEP A DIARY OR JOURNAL
S001605	HOW OFTEN DO YOU DO A CROSSWORD PUZZLE
S001606	HOW OFTEN DO YOU HELP OTHER STUDENTS WITH WRITING
S001607	HOW OFTEN DO YOU WRITE ABOUT WHAT YOU HAVE READ
S001608	HOW OFTEN DO YOU WRITE PAPERS TOO PERSONAL TO SHO
S001609	HOW OFTEN DO YOU WRITE FOR SCHOOL NEWSPAPER

Behavior around reading

S003501	HOW OFTEN DO YOU READ FOR FUN ON YOUR OWN TIME
S003502	HOW OFTEN DO YOU TELL FRIEND ABOUT A GOOD BOOK
S003503	HOW OFTEN DO YOU TAKE BOOKS OUT OF THE LIBRARY
S003504	HOW OFTEN DO YOU SPEND YOUR OWN MONEY ON BOOKS
S003505	HOW OFTEN DO YOU READ BOOK BASED ON MOVIE YOU SAW
S003506	HOW OFTEN DO YOU READ BOOKS BY AN AUTHOR YOU LIKE
S004401	HOW OFTEN DOES SOMEONE READ ALOUD TO YOU
S004402	HOW OFTEN DO YOU READ ALOUD TO SOMEONE
S005201	HOW OFTEN DO YOU READ ALOUD IN SCHOOL
S005202	HOW OFTEN DO YOU READ ON OWN IN SCHOOL
S005203	HOW OFTEN DO YOU WORK IN A WORKBOOK

Studying for tests

S005101	HOW OFTEN WHEN STUDY FOR TEST: READ OVER MATERIAL
S005102	HOW OFTEN WHEN STUDY FOR TEST: TAKE NOTES ON READ
S005103	HOW OFTEN WHEN STUDY FOR TEST: MAKE OUTLINES
S005104	HOW OFTEN WHEN STUDY FOR TEST: QUES IN TEXTBOOK
S005105	HOW OFTEN WHEN STUDY FOR TEST: ANSWER OWN QUESTNS
S005106	HOW OFTEN WHEN STUDY FOR TEST: QUESTION OTHERS

Use of library

S005301	HOW OFTEN GO TO LIBRARY TO READ ON OWN

Table A(2)
(continued)

S005302	HOW OFTEN GO TO LIBRARY TO LOOK UP FACT FOR SCHOO
S005303	HOW OFTEN GO TO LIBRARY TO FIND BOOKS FOR HOBBIES
S005304	HOW OFTEN GO TO LIBRARY FOR QUIET PLACE TO READ
S005305	HOW OFTEN GO TO LIBRARY TO TAKE OUT BOOKS

Behavior around writing in school

S002001	WHAT WAS THE LAST THING YOU WROTE IN SCHOOL
S002002	LAST WRITING IN SCHOOL: COPY OVER BEFORE SUBMITIN
S002003	LAST WRITING IN SCHOOL: MAKE CHANGES BEFORE SUBMI
S002004	LAST WRITING IN SCHOOL: MAKE CHANGES AFTER RETURN
S002005	LAST WRITING IN SCHOOL: LIKE DOING THE WRITING

Student orientation toward usefulness of reading and writing

Student orientation toward usefulness of writing

S000701	HOW OFTEN IS TRUE: WRITING IS IMPORTANT
S000702	HOW OFTEN IS TRUE: WRITING HELPS LEARN MYSELF
S000703	HOW OFTEN IS TRUE: WRITING REMINDS ABOUT THINGS
S000704	HOW OFTEN IS TRUE: WRITING HELPS ME STUDY
S000705	HOW OFTEN IS TRUE: WRITING HELPS NEW IDEAS
S001401	HOW OFTEN IS TRUE: GOOD WRITING GETS A BETTER JOB
S001402	HOW OFTEN TRUE: GOOD WRITING INFLUENTIAL
S001501	HOW OFTEN TRUE: WRITING HELPS THINK MORE CLEARLY
S001502	HOW OFTEN TRUE: WRITING HELPS TELL OTHERS THINKIN
S001503	HOW OFTEN TRUE: WRITING HELPS TELL OTHERS FEELING
S001504	HOW OFTEN TRUE: WRITING HELPS UNDERSTAND MYSELF

Student orientation toward usefulness of reading

S004201	HOW OFTEN READING: HELPS ME DECIDE WANT TO BE
S004202	HOW OFTEN READING: HELP ME LEARN TO FIX THINGS
S004203	HOW OFTEN READING: HELPS UNDERSTAND PEOPLES ACTIO
S004204	HOW OFTEN READING: READING IS IMPORTANT
S004205	HOW OFTEN READING: BETTER FEWER HARD WORDS
S004206	HOW OFTEN READING: BETTER FEWER LONG SENTENCES
S004207	HOW OFTEN READING: BETTER IF IT MATTERED TO ME
S004208	HOW OFTEN READING: BETTER IF TEACH GAVE MORE TIME
S004209	HOW OFTEN READING: BETTER IF DIDNT HAVE SO MUCH
S004210	HOW OFTEN READING: BETTER IF WASNT TESTED ON IT
S004211	HOW OFTEN READING: LIKE MORE IF COULD TALK W OTHE
S004801	HOW OFTEN TRUE: WRITING HELPS ME GET A GOOD JOB
S004802	HOW OFTEN TRUE: WRITING HELPS ME SHARE MY IDEAS
S004803	HOW OFTEN TRUE: WRITING HELPS SHOW I KNOW THINGS
S004804	HOW OFTEN TRUE: WRITING HELPS KEEP IN TOUCH FRIEN

Table A(2)
(continued)

Student's experiential base for writing

Student's experiential base for writing

S005401	HOW OFTEN DO YOU WATCH NEWS ON TELEVISION
S005402	HOW OFTEN DO YOU READ A NEWS MAGAZINE
S005403	HOW OFTEN DO YOU READ NEWSPAPER NOT COMICS OR SPR
S005404	HOW OFTEN DO YOU LISTEN TO NEWS ON RADIO

Reading and writing behavior of people in student's home

Reading behavior of people in student's home

S004501	HOW OFTEN DOES FAMILY READ NEWSPAPERS
S004502	HOW OFTEN DOES FAMILY READ MAGAZINES
S004503	HOW OFTEN DOES FAMILY READ BOOKS
S004504	HOW OFTEN DOES FAMILY READ RECIPES

Writing behavior of people in student's home

S001101	HOW OFTEN DOES FAMILY LIST THINGS TO DO
S001102	HOW OFTEN DOES FAMILY COPY RECIPES OR DIRECTIONS
S001103	HOW OFTEN DOES FAMILY FILL OUT ORDER BLANKS
S001104	HOW OFTEN DOES FAMILY WRITE CHECKS
S001105	HOW OFTEN DOES FAMILY KEEP DIARIES
S001106	HOW OFTEN DOES FAMILY WORK CROSSWORD PUZZLE
S001801	HOW OFTEN DOES FAMILY WRITE LETTER TO A RELATIVE
S001802	HOW OFTEN DOES FAMILY WRITE NOTES OR MESSAGES
S001803	HOW OFTEN DOES FAMILY WRITE STORY OR POEM
S001804	HOW OFTEN DOES FAMILY WRITE BUSINESS LETTER

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
1.	B000101	ETHNICITY	CB-01	CB-01	CB-01
2.	B000201	ARE YOU HISPANIC	CB-02	CB-02	CB-02
3.	B000301	WHAT LANGUAGE DO YOU SPEAK MOST OFTEN IN HOME	CB-03	CB-03	CB-03
4.	B000302	OTHER LANGUAGE YOU SPEAK MOST OFTEN IN HOME	CB-03	CB-03	CB-03
5.	B000401	WHAT LANGUAGE DO OTHERS SPEAK MOST OFTEN IN HOME	CB-04	CB-04	CB-04
6.	B000402	OTHER LANGUAGE OTHERS SPEAK MOST OFTEN IN HOME	CB-04	CB-04	CB-04
7.	B000501	FIRST OTHER LANGUAGE YOU KNOW	CB-05	CB-05	CB-05
8.	B000502	SECOND OTHER LANGUAGE YOU KNOW	CB-05	CB-05	CB-05
9.	B000503	THIRD OTHER LANGUAGE YOU KNOW	CB-05	CB-05	CB-05
10.	B000601	HOW FAR IN SCHOOL DID YOUR FATHER GO	CB-08	CB-08	CB-08
11.	B000701	HOW FAR IN SCHOOL DID YOUR MOTHER GO	CB-07	CB-07	CB-07
12.	B000801	DOES YOUR MOTHER WORK OUTSIDE YOUR HOME	CB-06	CB-06	CB-06
13.	B000901	DOES YOUR FAMILY GET A NEWSPAPER REGULARLY	CB-09	CB-09	CB-09
14.	B000902	IS THERE A DICTIONARY IN YOUR HOME	CB-10	CB-10	CB-10
15.	B000903	IS THERE AN ENCYCLOPEDIA IN YOUR HOME	CB-11	CB-11	CB-11
16.	B000904	ARE THERE MORE THAN 25 BOOKS IN YOUR HOME	CB-12	CB-12	CB-12
17.	B000905	DOES YOUR FAMILY GET MAGAZINES REGULARLY	CB-13	CB-13	CB-13
18.	B000906	IS THERE A VIDEO GAME IN YOUR HOME	CB-14	CB-14	CB-14
19.	B000907	IS THERE A COMPUTER IN YOUR HOME	CB-15	CB-15	CB-15
20.	B001001	DO YOU HAVE GYM ONCE PER WEEK	CB-17	CB-17	CB-17
21.	B001002	DO YOU HAVE ART ONCE PER WEEK	CB-18	CB-18	CB-18
22.	B001003	DO YOU HAVE MUSIC ONCE PER WEEK	CB-19	CB-19	CB-19
23.	B001004	DO YOU HAVE FOREIGN LANGUAGE ONCE PER WEEK	CB-20	CB-20	CB-20
24.	B001005	DO YOU HAVE COMPUTER CLASS ONCE PER WEEK	CB-21	CB-21	CB-21
25.	B001006	DO YOU HAVE DRAMA CLASS ONCE PER WEEK	CB-22	CB-22	CB-22
26.	B001007	DO YOU HAVE SCIENCE ONCE PER WEEK	CB-23	CB-23	CB-23
27.	B001101	HOW MANY PAGES READ IN SCHOOL AND FOR HOMEWORK	CB-24	CB-24	CB-24
28.	B001201	STORIES WRITTEN FOR ENGLISH LAST WEEK	CB-25	CB-25	CB-25
29.	B001202	ESSAYS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-26	CB-26	CB-26
30.	B001203	POEMS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-27	CB-27	CB-27
31.	B001204	PLAYS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-28	CB-28	CB-28
32.	B001205	LETTERS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-29	CB-29	CB-29
33.	B001206	BOOK REPORTS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-30	CB-30	CB-30
34.	B001207	OTHER REPORTS WRITTEN FOR ENGLISH CLASS LAST WEEK	CB-31	CB-31	CB-31
35.	B001208	I DO NOT HAVE AN ENGLISH CLASS	CB-32	CB-32	CB-32
36.	B001401	WHAT WAS THE LAST THING YOU READ FOR SCHOOL	CB-34	CB-34	CB-34
37.	B001501	WHAT WAS THE LAST THING YOU READ ON YOUR OWN	CB-35	CB-35	CB-35
38.	B001701	HOW MUCH TIME DID YOU SPEND ON HOMEWORK YESTERDAY	CB-33	CB-33	CB-33
39.	B001801	HOW MUCH TELEVISION DO YOU WATCH EACH DAY	CB-37	CB-37	CB-37
40.	B001901	WHICH DESCRIBES YOUR GRADES IN SCHOOL	CB-36	CB-36	CB-36
41.	B002001	HOW MANY DIFFERENT TOWNS HAVE YOU LIVED IN	CB-16	CB-16	CB-16
42.	B002701	COURSE WORK COMPLETED: MATHEMATICS			CB-38
43.	B002702	COURSE WORK COMPLETED: ENGLISH OR LITERATURE			CB-39
44.	B002703	COURSE WORK COMPLETED: JOURNALISM			CB-40
45.	B002704	COURSE WORK COMPLETED: FOREIGN LANGUAGE			CB-41

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
46.	B002705	COURSE WORK COMPLETED: HISTORY OR SOCIAL STUDIES			CB-42
47.	B002706	COURSE WORK COMPLETED: SCIENCE			CB-43
48.	B002707	COURSE WORK COMPLETED: COMPUTERS OR PROGRAMMING			CB-44
49.	B002708	COURSE WORK COMPLETED: BUSINESS OR VOCATIONAL			CB-45
50.	B002709	COURSE WORK COMPLETED: ARTS			CB-46
51.	B002710	COURSE WORK COMPLETED: MUSIC			CB-47
52.	B002801	HOURS PER WEEK WORKING IN PART-TIME JOB			CB-48
53.	S000101	TIME SPENT IN ENGLISH CLASS LEARNING TO WRITE	A-01 U-08	A-01 U-08	A-01 U-08
54.	S000201	REPORTS AND PAPERS WRITTEN FOR SCHOOL LAST 6 WEEKS	A-02 U-09	A-02 U-09	A-02 U-09
55.	S000301	WRITINGS DONE LAST WEEK FOR SOCIAL STUDIES CLASS	A-03 U-10	A-03 U-10	A-03 U-10
56.	S000401	WRITINGS DONE LAST WEEK FOR SCIENCE	A-04 U-11	A-04 U-11	A-04 U-11
57.	S000501	WRITINGS DONE LAST WEEK NON-SCHOOL RELATED	A-05	A-05	A-05
58.	S000601	WHEN WRITING HOW OFTEN TEACHER ASKS TO MAKE NOTES	A-06 U-01	A-06 U-01	A-06 U-01
59.	S000602	WHEN WRITING HOW OFTEN TEACHER ASKS MAKE OUTLINE	A-07 U-02	A-07 U-02	A-07 U-02
60.	S000603	WHEN WRITING HOW OFTEN TEACHER ASKS NOTE CHANGES	A-08 U-03	A-08 U-03	A-08 U-03
61.	S000604	WHEN WRITING HOW OFTEN TEACHER ASKS TALK TEACHER	A-09 U-04	A-09 U-04	A-09 U-04
62.	S000605	WHEN WRITING HOW OFTEN TEACHER ASKS TALK MATES	A-10 U-05	A-10 U-05	A-10 U-05
63.	S000606	WHEN WRITING HOW OFTEN TEACHER ASK REDO BEFOR GRD	A-11 U-06	A-11 U-06	A-11 U-06
64.	S000607	WHEN WRITING HOW OFTEN TEACHER ASK REDO AFTER GRD	A-12 U-07	A-12 U-07	A-12 U-07
65.	S000701	HOW OFTEN IS TRUE: WRITING IS IMPORTANT	B-01	B-01	B-01
66.	S000702	HOW OFTEN IS TRUE: WRITING HELPS LEARN ABOUT SELF	B-02	B-02	B-02
67.	S000703	HOW OFTEN IS TRUE: WRITING REMINDS ABOUT THINGS	B-03	B-03	B-03
68.	S000704	HOW OFTEN IS TRUE: WRITING HELPS ME STUDY	B-04	B-04	B-04
69.	S000705	HOW OFTEN IS TRUE: WRITING HELPS NEW IDEAS	B-05	B-05	B-05
70.	S000901	WHEN WRITING HOW OFTEN ASK SELF SUBJECT PAPER	B-06 V-01	B-06 V-01	B-06 V-01
71.	S000902	WHEN WRITING HOW OFTEN LOOK UP FACTS IN BOOKS	B-07 V-02	B-07 V-02	B-07 V-02
72.	S000903	WHEN WRITING HOW OFTEN THINK BEFORE WRITING	B-08 V-03	B-08 V-03	B-08 V-03
73.	S000904	WHEN WRITING HOW OFTEN THINK ABOUT ORGANIZATION	B-09 V-04	B-09 V-04	B-09 V-04
74.	S000905	WHEN WRITING HOW OFTEN USE DIFF STYLES PER PERSON	B-10 V-05	B-10 V-05	B-10 V-05
75.	S000906	WHEN WRITING HOW OFTEN MAKE CHANGES AS YOU WRITE	B-11 V-06	B-11 V-06	B-11 V-06
76.	S000907	WHEN WRITING HOW OFTEN MAKE CHANGES AFTER WRITING	B-12 V-07	B-12 V-07	B-12 V-07
77.	S001001	HOW OFTEN HAVE YOU SHOWN FRIENDS YOUR WRITINGS	B-13 Q-07	B-13 Y-01	B-13 Y-01
78.	S001002	HOW OFTEN HAVE PAPERS BEEN PRINTED IN SCHOOL PAPER	B-14 Q-08	B-14 Y-02	B-14 Y-02
79.	S001003	HOW OFTEN DOES YOUR FAMILY READ YOUR PAPERS	B-15 Q-09	B-15 Y-03	B-15 Y-03
80.	S001101	HOW OFTEN DOES FAMILY LIST THINGS TO BUY OR DO	C-01	C-01	C-01
81.	S001102	HOW OFTEN DOES FAMILY COPY RECIPES OR DIRECTIONS	C-02	C-02	C-02
82.	S001103	HOW OFTEN DOES FAMILY FILL OUT ORDER BLANKS	C-03	C-03	C-03
83.	S001104	HOW OFTEN DOES FAMILY WRITE CHECKS/KEEP BUDGETS	C-04	C-04	C-04
84.	S001105	HOW OFTEN DOES FAMILY KEEP DIARIES OR JOURNALS	C-05	C-05	C-05
85.	S001106	HOW OFTEN DOES FAMILY WORK CROSSWORD PUZZLE	C-06	C-06	C-06
86.	S001201	HOW OFTEN IS TRUE: I LIKE TO WRITE	C-07	C-07	C-07
87.	S001202	HOW OFTEN IS TRUE: I AM A GOOD WRITER	C-08	C-08	C-08
88.	S001203	HOW OFTEN IS TRUE: THINK WRITING IS WASTE OF TIME	C-09	C-09	C-09
89.	S001204	HOW OFTEN IS TRUE: PEOPLE LIKE WHAT I WRITE	C-10	C-10	C-10
90.	S001205	HOW OFTEN IS TRUE: WRITE ON OWN AWAY FROM SCHOOL	C-11	C-11	C-11

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
91.	S001206	HOW OFTEN IS TRUE: DISLIKE WRITING TO BE GRADED	C-12	C-12	C-12
92.	S001207	HOW OFTEN IS TRUE: WOULDNT WRITE IF NOT FOR SCHOOL	C-13	C-13	C-13
93.	S001301	HOW OFTEN IS TRUE: MOVE SENTENCES AROUND	C-14 X-01	C-14 X-01	C-14 X-01
94.	S001302	HOW OFTEN IS TRUE: ADD NEW IDEAS OR INFORMATION	C-15 X-02	C-15 X-02	C-15 X-02
95.	S001303	HOW OFTEN IS TRUE: TAKE OUT UNDESIRED PARTS	C-16 X-03	C-16 X-03	C-16 X-03
96.	S001304	HOW OFTEN IS TRUE: CHANGE WORDS	C-17 X-04	C-17 X-04	C-17 X-04
97.	S001305	HOW OFTEN IS TRUE: CORRECT SPELLING MISTAKES	C-18 X-05	C-18 X-05	C-18 X-05
98.	S001306	HOW OFTEN IS TRUE: CORRECT GRAMMAR MISTAKES	C-19 X-06	C-19 X-06	C-19 X-06
99.	S001307	HOW OFTEN IS TRUE: CORRECT PUNCTUATION MISTAKES	C-20 X-07	C-20 X-07	C-20 X-07
100.	S001308	HOW OFTEN IS TRUE: REWRITE MOST OF PAPER	C-21 X-08	C-21 X-08	C-21 X-08
101.	S001309	HOW OFTEN IS TRUE: THROW OUT AND START OVER	C-22 X-09	C-22 X-09	C-22 X-09
102.	S001401	HOW OFTEN IS TRUE: GOOD WRITING GETS A BETTER JOB	D-01	D-01	D-01
103.	S001402	HOW OFTEN IS TRUE: GOOD WRITING MORE INFLUENTIAL	D-02	D-02	D-02
104.	S001501	HOW OFTEN TRUE: WRITING HELPS THINK MORE CLEARLY	D-03	D-03	D-03
105.	S001502	HOW OFTEN TRUE: WRITING HELPS TELL OTHERS THINKING	D-04	D-04	D-04
106.	S001503	HOW OFTEN TRUE: WRITING HELPS TELL OTHERS FEELINGS	D-05	D-05	D-05
107.	S001504	HOW OFTEN TRUE: WRITING HELPS UNDERSTAND MYSELF	D-06	D-06	D-06
108.	S001601	HOW OFTEN DO YOU LIST THINGS TO BUY	D-07	D-07	D-07
109.	S001602	HOW OFTEN DO YOU COPY RECIPES OR DIRECTIONS	D-08	D-08	D-08
110.	S001603	HOW OFTEN DO YOU FILL OUT ORDER BLANKS	D-09	D-09	D-09
111.	S001604	HOW OFTEN DO YOU KEEP A DIARY OR JOURNAL	D-10	D-10	D-10
112.	S001605	HOW OFTEN DO YOU DO A CROSSWORD PUZZLE	D-11	D-11	D-11
113.	S001606	HOW OFTEN DO YOU HELP OTHER STUDENTS WITH WRITING	D-12	D-12	D-12
114.	S001607	HOW OFTEN DO YOU WRITE ABOUT WHAT YOU HAVE READ	D-13	D-13	D-13
115.	S001608	HOW OFTEN DO YOU WRITE PAPERS TOO PERSONAL TO SHOW	D-14	D-14	D-14
116.	S001609	HOW OFTEN DO YOU WRITE FOR SCHOOL NEWSPAPER	D-15	D-15	D-15
117.	S001701	HOW OFTEN DOES TEACHER TALK RE: FOLLOW DIRECTIONS	D-16 W-23	D-16 W-23	D-16 W-05
118.	S001702	HOW OFTEN DOES TEACHER TALK RE: WROTE ENOUGH	D-17 W-24	D-17 W-24	D-17 W-06
119.	S001703	HOW OFTEN DOES TEACHER TALK RE: IDEAS IN PAPER	D-18 W-25	D-18 W-25	D-18 W-07
120.	S001704	HOW OFTEN DOES TEACHER TALK RE: EXPLAIN IN PAPER	D-19 W-26	D-19 W-26	D-19 W-08
121.	S001705	HOW OFTEN DOES TEACHER TALK RE: FEELINGS IN PAPER	D-20 W-27	D-20 W-27	D-20 W-09
122.	S001706	HOW OFTEN DOES TEACHER TALK RE: ORGANIZING PAPER	D-21 W-28	D-21 W-28	D-21 W-10
123.	S001707	HOW OFTEN DOES TEACHER TALK RE: WORDS IN PAPER	D-22 W-29	D-22 W-29	D-22 W-11
124.	S001708	HOW OFTEN DOES TEACHER TALK RE: SP, GRAM IN PAPER	D-23 W-30	D-23 W-30	D-23 W-12
125.	S001709	HOW OFTEN DOES TEACHER TALK RE: NEATNESS IN PAPER	D-24 W-31	D-24 W-31	D-24 W-13
126.	S001801	HOW OFTEN DOES FAMILY WRITE LETTERS TO RELATIVES	E-01	E-01	E-01
127.	S001802	HOW OFTEN DOES FAMILY WRITE NOTES OR MESSAGES	E-02	E-02	E-02
128.	S001803	HOW OFTEN DOES FAMILY WRITE STORIES OR POEMS	E-03	E-03	E-03
129.	S001804	HOW OFTEN DOES FAMILY WRITE BUSINESS LETTERS	E-04	E-04	E-04
130.	S001901	HOW OFTEN DO YOU WRITE A BOOK REPORT	E-05	E-05	E-05
131.	S001902	HOW OFTEN DO YOU WRITE ABOUT SCIENCE EXPERIMENT	E-06	E-06	E-06
132.	S001903	HOW OFTEN DO YOU WRITE LETTER TO A RELATIVE	E-07	E-07	E-07
133.	S001904	HOW OFTEN DO YOU WRITE NOTES OR MESSAGES	E-08	E-08	E-08
134.	S001905	HOW OFTEN DO YOU WRITE STORIES THAT NOT HOMEWORK	E-09	E-09	E-09
135.	S002001	WHAT WAS THE LAST THING YOU WROTE IN SCHOOL	F-01 V-14	F-01 V-14	F-01 V-14

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
136.	S002002	LAST WRITING IN SCHOOL: COPY OVER BEFORE SUBMITING	F-02 V-15	F-02 V-15	F-02 V-15
137.	S002003	LAST WRITING IN SCHOOL: MAKE CHANGES BEFORE SUBMIT	F-03 V-16	F-03 V-16	F-03 V-16
138.	S002004	LAST WRITING IN SCHOOL: MAKE CHANGES AFTER RETURND	F-04 V-17	F-04 V-17	F-04 V-17
139.	S002005	LAST WRITING IN SCHOOL: LIKE DOING THE WRITING	F-05 V-18	F-05 V-18	F-05 V-18
140.	S002501	HOW OFTEN DOES TEACHER MARK ERRORS ON PAPERS	G-01 V-19	G-01 V-19	G-01 V-19
141.	S002502	HOW OFTEN DOES TEACHER WRITE NOTES ON PAPERS	G-02 V-20	G-02 V-20	G-02 V-20
142.	S002503	HOW OFTEN DOES TEACHER POINT OUT GOOD THINGS	G-03 V-21	G-03 V-21	G-03 V-21
143.	S002504	HOW OFTEN DOES TEACHER POINT OUT NOT GOOD THINGS	G-04 V-22	G-04 V-22	G-04 V-22
144.	S002505	HOW OFTEN DOES TEACHER MAKE SUGGESTIONS FOR NEXT	G-05 V-23	G-05 V-23	G-05 V-23
145.	S002506	HOW OFTEN DOES TEACHER SHOW INTEREST IN WRITING	G-06 V-24	G-06 V-24	G-06 V-24
146.	S002701	DID YOU GO TO KINDERGARTEN	H-01	H-01	H-01
147.	S002702	DID YOU GO TO DAY CARE	H-02	H-02	H-02
148.	S002703	DID YOU GO TO NURSERY SCHOOL	H-03	H-03	H-03
149.	S002704	DID YOU GO TO HEADSTART	H-04	H-04	H-04
150.	S002801	WHERE DID YOU LIVE AT AGE 9		H-05	H-05
151.	S002802	WHERE DID YOU LIVE AT AGE 9: STATE		H-05	H-05
152.	S002803	WHERE DID YOU LIVE AT AGE 9: COUNTRY		H-05	H-05
153.	S002804	RESIDENCE AT AGE 9 VS. CURRENT RESIDENCE		H-05	H-05
154.	S002901	DO YOU USE A COMPUTER AT HOME	J-01	J-01	J-01
155.	S002902	DO YOU USE A COMPUTER AT THE LIBRARY	J-02	J-02	J-02
156.	S002903	DO YOU USE A COMPUTER AT A FRIENDS HOUSE	J-03	J-03	J-03
157.	S002904	HOW OFTEN DO YOU USE A COMPUTER AT SCHOOL	J-04	J-04	J-04
158.	S003001	DO YOU USE A COMPUTER TO PLAY GAMES	J-05	J-05	J-05
159.	S003002	DO YOU USE A COMPUTER TO LEARN THINGS	J-06	J-06	J-06
160.	S003003	DO YOU USE A COMPUTER TO WRITE STORIES OR PAPERS	J-07	J-07	J-07
161.	S003101	HOW OFTEN DO YOU WRITE COMPUTER PROGRAMS	J-08	J-08	J-08
162.	S003201	WHAT DO YOU USUALLY DO AFTER SCHOOL	J-09	J-09	
163.	S003202	IF YOU GO HOME AFTER SCHOOL, WHO IS USUALLY THERE	J-10	J-10	
164.	S003203	WHAT OTHER ADULT	J-10	J-10	
165.	S003301	WHAT KIND OF READER ARE YOU	K-01 V-28	K-01 V-28	K-01
166.	S003401	DO YOU EXPECT TO GRADUATE FROM HIGH SCHOOL	K-02	K-02	K-02
167.	S003501	HOW OFTEN DO YOU READ FOR FUN ON YOUR OWN TIME	K-03 V-08	K-03 V-08	K-03 V-08
168.	S003502	HOW OFTEN DO YOU TELL A FRIEND ABOUT A GOOD BOOK	K-04 V-09	K-04 V-09	K-04 V-09
169.	S003503	HOW OFTEN DO YOU TAKE BOOKS OUT OF THE LIBRARY	K-05 V-10	K-05 V-10	K-05 V-10
170.	S003504	HOW OFTEN DO YOU SPEND YOUR OWN MONEY ON BOOKS	K-06 V-11	K-06 V-11	K-06 V-11
171.	S003505	HOW OFTEN DO YOU READ BOOK BASED ON MOVIE YOU SAW	K-07 V-12	K-07 V-12	K-07 V-12
172.	S003506	HOW OFTEN DO YOU READ BOOKS BY AN AUTHOR YOU LIKE	K-08 V-13	K-08 V-13	K-08 V-13
173.	S003601	HOW OFTEN DO YOU GO TO A MOVIE	L-01	L-01	L-01
174.	S003602	HOW OFTEN DO YOU GO TO A PLAY	L-02	L-02	L-02
175.	S003603	HOW OFTEN DO YOU GO TO A CONCERT	L-03	L-03	L-03
176.	S003604	HOW OFTEN DO YOU GO TO A PARTY	L-04	L-04	L-04
177.	S003605	HOW OFTEN DO YOU GO TO THE PUBLIC LIBRARY	L-05	L-05	L-05
178.	S003606	HOW OFTEN DO YOU TRAVEL TO A PLACE AWAY FROM HOME	L-06	L-06	L-06
179.	S003607	HOW OFTEN DO YOU GO SHOPPING	L-07	L-07	L-07
180.	S003608	HOW OFTEN DO YOU GO TO A SPORTS EVENT	L-08	L-08	L-08

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
181.	S003609	HOW OFTEN DO YOU PLAY CARD OR TABLE GAMES	L-09	L-09	L-09
182.	S003610	HOW OFTEN DO YOU VISIT RELATIVES	L-10	L-10	L-10
183.	S003611	HOW OFTEN DO YOU GO TO A MUSEUM	L-11	L-11	L-11
184.	S003612	HOW OFTEN DO YOU GO CAMPING	L-12	L-12	L-12
185.	S003613	HOW OFTEN DO YOU STAY HOME ALONE	L-13	L-13	L-13
186.	S003614	WHAT ACTIVITY DO YOU DO MOST OFTEN	L-14	L-14	L-14
187.	S003701	DO YOU EVER FEEL BORED AT SCHOOL	L-15	L-15	L-15
188.	S003801	DURING PAST YEAR HOW OFTEN SENT TO PRINCIPALS OFF	L-16	L-16	L-16
189.	S003802	DURING PAST YEAR HOW OFTEN PLACED ON PROBATION		L-17	L-17
190.	S003803	DURING PAST YEAR HOW OFTEN GIVEN DETENTION		L-18	L-18
191.	S003804	DURING PAST YEAR HOW OFTEN WARNED ABOUT ATTENDANCE	L-17	L-19	L-19
192.	S003805	DURING PAST YEAR HOW OFTEN WARNED ABOUT GRADES	L-18	L-20	L-20
193.	S003806	DURING PAST YEAR HOW OFTEN WARNED ABOUT BEHAVIOR	L-19	L-21	L-21
194.	S003901	HOW MANY OLDER BROTHERS AND SISTERS	M-01	M-01	M-01
195.	S003902	HOW MANY YOUNGER BROTHERS AND SISTERS	M-02	M-02	M-02
196.	S004001	HOW MANY DAYS OF SCHOOL MISSED LAST MONTH	M-03	M-03	M-03
197.	S004101	HOW MANY TIMES LATE FOR SCHOOL LAST MONTH	M-04	M-04	M-04
198.	S004201	HOW OFTEN READING: HELPS ME DECIDE WANT TO BE	N-01	N-01	N-01
199.	S004202	HOW OFTEN READING: HELPS ME LEARN TO FIX THINGS	N-02	N-02	N-02
200.	S004203	HOW OFTEN READING: HELPS UNDERSTAND PEOPLES ACTION	N-03	N-03	N-03
201.	S004204	HOW OFTEN READING: READING IS IMPORTANT	N-04	N-04	N-04
202.	S004205	HOW OFTEN READING: BETTER FEWER HARD WORDS	N-05	N-05	N-05
203.	S004206	HOW OFTEN READING: BETTER FEWER LONG SENTENCES	N-06	N-06	N-06
204.	S004207	HOW OFTEN READING: BETTER IF IT MATTERED TO ME	N-07	N-07	N-07
205.	S004208	HOW OFTEN READING: BETTER IF TEACH GAVE MORE TIME	N-08	N-08	N-08
206.	S004209	HOW OFTEN READING: BETTER IF DIDNT HAVE SO MUCH	N-09	N-09	N-09
207.	S004210	HOW OFTEN READING: BETTER IF WASNT TESTED ON IT	N-10	N-10	N-10
208.	S004211	HOW OFTEN READING: LIKE MORE IF COULD TALK W OTHER	N-11	N-11	N-11
209.	S004301	HOW OFTEN DO YOU READ A STORY OR NOVEL	O-01	O-01	O-01
210.	S004302	HOW OFTEN DO YOU READ A POEM	O-02	O-02	O-02
211.	S004303	HOW OFTEN DO YOU READ A PLAY	O-03	O-03	O-03
212.	S004304	HOW OFTEN DO YOU READ A NEWSPAPER	O-04	O-04	O-04
213.	S004305	HOW OFTEN DO YOU READ A MAGAZINE	O-05	O-05	O-05
214.	S004306	HOW OFTEN DO YOU READ A SCIENCE BOOK	O-06	O-06	O-06
215.	S004307	HOW OFTEN DO YOU READ A BIOGRAPHY	O-07	O-07	O-07
216.	S004308	HOW OFTEN DO YOU READ A HOW-TO-DO BOOK	O-08	O-08	O-08
217.	S004309	HOW OFTEN DO YOU READ A BOOK ABOUT OTHER TIMES	O-09	O-09	O-09
218.	S004310	HOW OFTEN DO YOU READ A SPORTS BOOK	O-10	O-10	O-10
219.	S004311	HOW OFTEN DO YOU READ WORDS OF A SONG	O-11	O-11	O-11
220.	S004401	HOW OFTEN DOES SOMEONE READ ALOUD TO YOU	P-01	P-01	P-01
221.	S004402	HOW OFTEN DO YOU READ ALOUD TO SOMEONE	P-02	P-02	P-02
222.	S004501	HOW OFTEN DOES FAMILY READ NEWSPAPERS	P-03	P-03	P-03
223.	S004502	HOW OFTEN DOES FAMILY READ MAGAZINES	P-04	P-04	P-04
224.	S004503	HOW OFTEN DOES FAMILY READ BOOKS	P-05	P-05	P-05
225.	S004504	HOW OFTEN DOES FAMILY READ RECIPES	P-06	P-06	P-06

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
226.	S004601	HOW OFTEN WITH NEW READING TEACHER POINT HARD WORD	Q-01 U-12 V-25 X-10	Q-01 U-12 V-25 X-10	Q-01 U-12 V-25 X-10
227.	S004602	HOW OFTEN WITH NEW READING TEACHER PREVIEW READING	Q-02 U-13 V-26 X-11	Q-02 U-13 V-26 X-11	Q-02 U-13 V-26 X-11
228.	S004603	HOW OFTEN WITH NEW READING TEACHER READ PART ALOUD	Q-03 U-14 V-27 X-12	Q-03 U-14 V-27 X-12	Q-03 U-14 V-27 X-12
229.	S004701	HOW OFTEN DOES TEACHER LIST OF QUESTS AS YOU READ	Q-04 U-15 X-13	Q-04 U-15 X-13	Q-04 U-15 X-13
230.	S004702	HOW OFTEN DOES TEACHER TELL HOW TO FIND MAIN IDEA	Q-05 U-16 X-14	Q-05 U-16 X-14	Q-05 U-16 X-14
231.	S004703	HOW OFTEN DOES TEACHER TELL HOW TO READ FASTER	Q-06 U-17 X-15	Q-06 U-17 X-15	Q-06 U-17 X-15
232.	S004801	HOW OFTEN TRUE: WRITING HELPS ME GET A GOOD JOB	R-01	R-01	R-01
233.	S004802	HOW OFTEN TRUE: WRITING HELPS ME SHARE MY IDEAS	R-02	R-02	R-02
234.	S004803	HOW OFTEN TRUE: WRITING HELPS SHOW I KNOW THINGS	R-03	R-03	R-03
235.	S004804	HOW OFTEN TRUE: WRITING HELPS KEEP IN TOUCH FRIEN	R-04	R-04	R-04
236.	S005001	WHEN FREE TIME, HOW OFTEN WATCH TV	S-01 W-01	S-01 W-01	S-01
237.	S005002	WHEN FREE TIME, HOW OFTEN READ A BOOK	S-02 W-02	S-02 W-02	S-02
238.	S005003	WHEN FREE TIME, HOW OFTEN WRITE IN DIARY	S-03 W-03	S-03 W-03	S-03
239.	S005004	WHEN FREE TIME, HOW OFTEN CALL A FRIEND	S-04 W-04	S-04 W-04	S-04
240.	S005005	WHEN FREE TIME, HOW OFTEN BE WITH FRIENDS	S-05 W-05	S-05 W-05	S-05
241.	S005006	WHEN FREE TIME, HOW OFTEN GO SHOPPING	S-06 W-06	S-06 W-06	S-06
242.	S005007	WHEN FREE TIME, HOW OFTEN PLAY A SPORT	S-07 W-07	S-07 W-07	S-07
243.	S005008	WHEN FREE TIME, HOW OFTEN GO HUNTING OR FISHING	S-08 W-08	S-08 W-08	S-08
244.	S005009	WHEN FREE TIME, HOW OFTEN TAKE A WALK	S-09 W-09	S-09 W-09	S-09
245.	S005010	WHEN FREE TIME, HOW OFTEN WORK AT A COMPUTER	S-10 W-10	S-10 W-10	S-10
246.	S005011	WHEN FREE TIME, HOW OFTEN PLAY VIDEO GAMES	S-11 W-11	S-11 W-11	S-11
247.	S005012	WHEN FREE TIME, HOW OFTEN READ A NEWSPAPER	S-12 W-12	S-12 W-12	S-12
248.	S005013	WHEN FREE TIME, HOW OFTEN GET A SNACK	S-13 W-13	S-13 W-13	S-13
249.	S005014	WHEN FREE TIME, HOW OFTEN DO EXTRA HOMEWORK	S-14 W-14	S-14 W-14	S-14
250.	S005015	WHEN FREE TIME, HOW OFTEN WRITE A LETTER	S-15 W-15	S-15 W-15	S-15
251.	S005016	WHEN FREE TIME, HOW OFTEN LISTEN TO MUSIC	S-16 W-16	S-16 W-16	S-16
252.	S005017	WHEN FREE TIME, HOW OFTEN DO SOMETHING ELSE	S-17 W-17	S-17 W-17	S-17
253.	S005019	WHEN FREE TIME WHAT ACTIVITY SPEND MOST TIME	S-18 W-18	S-18 W-18	S-18
254.	S005101	HOW OFTEN WHEN STUDY FOR TEST: READ OVER MATERIAL	T-01	T-01	T-01
255.	S005102	HOW OFTEN WHEN STUDY FOR TEST: TAKE NOTES ON READ	T-02	T-02	T-02
256.	S005103	HOW OFTEN WHEN STUDY FOR TEST: MAKE OUTLINES	T-03	T-03	T-03
257.	S005104	HOW OFTEN WHEN STUDY FOR TEST: QUES IN TEXTBOOK	T-04	T-04	T-04
258.	S005105	HOW OFTEN WHEN STUDY FOR TEST: ANSWER OWN QUESTNS	T-05	T-05	T-05
259.	S005106	HOW OFTEN WHEN STUDY FOR TEST: QUESTION OTHERS	T-06	T-06	T-06
260.	S005201	HOW OFTEN DO YOU READ ALOUD IN SCHOOL	T-07	T-07	T-07
261.	S005202	HOW OFTEN DO YOU READ ON YOUR OWN IN SCHOOL	T-08	T-08	T-08
262.	S005203	HOW OFTEN DO YOU WORK IN A WORKBOOK	T-09	T-09	T-09
263.	S005301	HOW OFTEN GO TO LIBRARY TO READ ON OWN	T-10 W-32	T-10 W-32	T-10 W-14
264.	S005302	HOW OFTEN GO TO LIBRARY TO LOOK UP FACT FOR SCHOOL	T-11 W-33	T-11 W-33	T-11 W-15
265.	S005303	HOW OFTEN GO TO LIBRARY TO FIND BOOKS FOR HOBBIES	T-12 W-34	T-12 W-34	T-12 W-16
266.	S005304	HOW OFTEN GO TO LIBRARY FOR QUIET PLACE TO READ	T-13 W-35	T-13 W-35	T-13 W-17
267.	S005305	HOW OFTEN GO TO LIBRARY TO TAKE OUT BOOKS	T-14 W-36	T-14 W-36	T-14 W-18
268.	S005401	HOW OFTEN DO YOU WATCH NEWS ON TELEVISION	T-15 W-19	T-15 W-19	T-15 W-01
269.	S005402	HOW OFTEN DO YOU READ A NEWS MAGAZINE	T-16 W-20	T-16 W-20	T-16 W-02
270.	S005403	HOW OFTEN DO YOU READ NEWSPAPER NOT COMICS OR SPRT	T-17 W-21	T-17 W-21	T-17 W-03

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
271.	S005404	HOW OFTEN DO YOU LISTEN TO NEWS ON RADIO	T-18 W-22	T-18 W-22	T-18 W-04
272.	S005701	HAVE YOU APPLIED FOR ADMISSION TO A COLLEGE OR UNV			J-10
273.	S005702	HAVE YOU APPLIED TO A FOUR-YEAR COLLEGE			J-10
274.	S005703	HAVE YOU APPLIED TO A TWO-YEAR COLLEGE			J-10
275.	S005704	HAVE YOU APPLIED TO OTHER COLLEGE OR UNIVERSITY			J-10
276.	S005801	WHAT ARE YOUR LONG-TERM CAREER GOALS			J-11
277.	S005802	LONG-TERM CAREER GOAL CODE			J-11
278.	S005901	WHERE DID YOU LIVE AT AGE 13			H-06
279.	S005902	WHERE DID YOU LIVE AT AGE 13: STATE			H-06
280.	S005903	WHERE DID YOU LIVE AT AGE 13: COUNTRY			H-06
281.	S005904	RESIDENCE AT AGE 13 VS. CURRENT RESIDENCE			H-06
282.	S006001	WHICH COURSES HAVE YOU TAKEN:GENERAL SCIENCE			L-22
283.	S006002	WHICH COURSES HAVE YOU TAKEN:BIOLOGY			L-23
284.	S006003	WHICH COURSES HAVE YOU TAKEN:CHEMISTRY			L-24
285.	S006004	WHICH COURSES HAVE YOU TAKEN:PHYSICS			L-25
286.	S006005	WHICH COURSES HAVE YOU TAKEN:OTHER SCIENCE (1)			L-26
287.	S006007	WHICH COURSES HAVE YOU TAKEN:OTHER SCIENCE (2)			L-26
288.	S006009	WHICH COURSES HAVE YOU TAKEN:OTHER SCIENCE (3)			L-26
289.	S006101	WHICH COURSES HAVE YOU TAKEN:GENERAL MATH 1			N-12
290.	S006102	WHICH COURSES HAVE YOU TAKEN:GENERAL MATH 2			N-13
291.	S006103	WHICH COURSES HAVE YOU TAKEN:FIRST YEAR ALGEBRA			N-14
292.	S006104	WHICH COURSES HAVE YOU TAKEN:SECOND YEAR ALGEBRA			N-15
293.	S006105	WHICH COURSES HAVE YOU TAKEN:GEOMETRY			N-16
294.	S006106	WHICH COURSES HAVE YOU TAKEN:CALCULUS			N-17
295.	S006107	WHICH COURSES HAVE YOU TAKEN:OTHER MATH COURSE (1)			N-18
296.	S006109	WHICH COURSES HAVE YOU TAKEN:OTHER MATH COURSE (2)			N-19
297.	S006111	WHICH COURSES HAVE YOU TAKEN:OTHER MATH COURSE (3)			N-20
298.	S006201	RATE YOUR SCHOOL IN:PREPARING STUDENTS FOR COLLEGE			P-07
299.	S006202	RATE YOUR SCHOOL IN:PREPARING STUDENTS FOR CAREER			P-08
300.	S006203	RATE YOUR SCHOOL IN:PREPARING STUDENTS FOR LIFE			P-09
301.	S006204	RATE YOUR SCHOOL IN:VARIETY OF EXTRACUR ACTIVITIES			P-10
302.	S006205	RATE YOUR SCHOOL IN:QUALITY OF EXTRACUR ACTIVITES			P-11
303.	S006206	RATE YOUR SCHOOL IN:FACULTY INTEREST IN STUDENTS			P-12
304.	S006207	RATE YOUR SCHOOL IN:QUALITY OF FACULTY			P-13
305.	S006208	RATE YOUR SCHOOL IN:QUALITY OF STUDENT LIFE			P-14
306.	S006301	SCHOOL EXPERIENCES:I AM SATSFIED WITH MY EDUCATION			R-05
307.	S006302	SCHOOL EXPERIENCES:NOT LEARNG WHAT I NEED TO KNOW			R-06
308.	S006303	SCHOOL EXPERIENCES:HAVE HAD DISCIPLNE PRBS THIS YR			R-07
309.	S006304	SCHOOL EXPERIENCES:I AM INTERESTED IN SCHOOL			R-08
310.	S006305	SCHOOL EXPERIENCES:ONCE IN A WHILE I CUT CLASS			R-09
311.	S006306	SCHOOL EXPERIENCES:I DON'T FEEL SAFE AT THIS SCHL			R-10
312.	S006307	SCHOOL EXPERIENCES:WISH I COULD GO TO DIFF SCHOOL			R-11
313.	S006401	HAVE YOU EVER BEEN IN PROGRAM:REMEDIAL ENGLISH			V-28
314.	S006402	HAVE YOU EVER BEEN IN PROGRAM:REMEDIAL MATH			V-29
315.	S006403	HAVE YOU EVER BEEN IN PROGRAM:HONORS ENGLISH			V-30

Table A(3)
Background and Attitude Items and Locations (Spiral)

N	NAEP ID	DESCRIPTION	Grade 4/Age 9	Grade 8/Age 13	Grade 11/Age 17
316.	S006404	HAVE YOU EVER BEEN IN PROGRAM:HONORS MATHEMATICS			V-31
317.	S006405	HAVE YOU EVER BEEN IN PROGRAM:HONORS SCIENCE			V-32
318.	S006406	HAVE YOU EVER BEEN IN PROGRAM:BILINGUAL PROGRAM			V-33
319.	S006407	HAVE YOU EVER BEEN IN PROGRAM:FAMILY-LIFE,SEX ED			V-34
320.	S006408	HAVE YOU EVER BEEN IN PROGRAM:ALCOHL,DRUG-ABUSE ED			V-35
321.	S006409	HAVE YOU EVER BEEN IN PROGRAM:SPEC PHYSICAL PROGRM			V-36
322.	S006410	HAVE YOU EVER BEEN IN PROGRAM:SPEC SPEECH PROGRAM			V-37
323.	S006501	HAVE YOU TAKEN COURSES:AGRICULTURE,INCLD HORTICULT			W-19
324.	S006502	HAVE YOU TAKEN COURSES:AUTO MECHANICS			W-20
325.	S006503	HAVE YOU TAKEN COURSES:COMMERCIAL ARTS			W-21
326.	S006504	HAVE YOU TAKEN COURSES:COMPUTER PROGRAMMING			W-22
327.	S006505	HAVE YOU TAKEN COURSES:CONSTRUCTION,CARPENTRY TRDS			W-23
328.	S006506	HAVE YOU TAKEN COURSES:CONSTRUCTION TRADES:ELECTRL			W-24
329.	S006507	HAVE YOU TAKEN COURSES:CONTRUCTION TRADES:MASONRY			W-25
330.	S006508	HAVE YOU TAKEN COURSES:CONSTRUCTION TRADES:PLUMBNG			W-26
331.	S006509	HAVE YOU TAKEN COURSES:COSMETOLOGY,HAIRDRESSING			W-27
332.	S006510	HAVE YOU TAKEN COURSES:DRAFTING			W-28
333.	S006511	HAVE YOU TAKEN COURSES:ELECTRONICS			W-29
334.	S006512	HAVE YOU TAKEN COURSES:HOME EC,DIETETICS,CHILD CAR			W-30
335.	S006513	HAVE YOU TAKEN COURSES:MACHINE SHOP			W-31
336.	S006514	HAVE YOU TAKEN COURSES:MEDICAL OR DENTAL ASSISTNT			W-32
337.	S006515	HAVE YOU TAKEN COURSES:PRACTICAL NURSING			W-33
338.	S006516	HAVE YOU TAKEN COURSES:FOOD SERVICE OCCUPATIONS			W-34
339.	S006517	HAVE YOU TAKEN COURSES:SALES OR MERCHANDISING			W-35
340.	S006518	HAVE YOU TAKEN COURSES:SECRETARIAL,OFFICE WORK			W-36
341.	S006519	HAVE YOU TAKEN COURSES:WELDING			W-37
342.	S006601	WHAT TAKE MOST OF YOUR TIME YEAR AFTER HIGH SCHOOL			Y-04
343.	S006701	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:WORK			Y-05
344.	S006702	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:APPRENTICE			Y-05
345.	S006703	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:MILITARY			Y-05
346.	S006704	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:HOMEMAKER			Y-05
347.	S006705	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:VOC SCHOOL			Y-05
348.	S006706	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:COMM COLLEG			Y-05
349.	S006707	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:VOC COURSES			Y-05
350.	S006708	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:4-YR COLLEG			Y-05
351.	S006709	OTHER PLANS FOR YEAR AFTER HIGH SCHOOL:TRAVEL,NONE			Y-05
352.	S006801	HOW MUCH FREE TIME ON AVERAGE SCHOOL DAY	J-11		J-09

APPENDIX B

Reading Trend Analysis Items

Table B-1

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	1	7099001-A001/001	N004002	TRIANGLE:NAME FIGURE AS TRIANGLE
R	2	7099001-A002/003	N004001	TRIANGLE:DRAWING TRIANGLE
R	3	7099002-A001/001	N001801	FLY:WANT OF THOUGHT-LACK OF THINKING
R	4	7099002-A002/002	N001802	FLY:FACING PROBLEMS SIMILAR TO HIS OWN
R	5	7099003-A001/001		WAYFARER:FEW SEEK TRUTH
R	6	7099004-A001/001		DROPOUT:DROPOUTS HAVE HARD TIME GETTING J
R	7	7099005-A001/001		ADM DRAKE:SENT PENGUIN
R	8	7099005-A002/002		PENGUIN CAPT COOKS HOME
R	9	7099005-A003/003		PENGUINS DIFFICULT PETS
R	10	7099006-A001/001	N004201	MEOW-WOW:2 MONTH KITTEN-FEED 3 OR 4 TMS D
R	11	7099006-A002/002	N004202	MEOW-WOW:CAT LEAVES FOOD-LEAVE BOWL FOR H
R	12	7099007-A001/001	N005001	ARTS:BEFORE 1940 ARTS WERE ORIENTED TO EL
R	13	7099007-A002/002	N005002	ARTS:PRIVILEGE OF ARISTOCRATIC FEW-GREAT
R	14	7099007-A003/003	N005003	ARTS:MASS PROD NO HARM TO GENUINE ART
R	15	7099008-A001/001		REASONS FOR DOG OVER CAT
R	16	7099009-A001/001	N003601	MAGIC TRICK:FIRST TIE BLACK THREAD
R	17	7099009-A002/002	N003602	MAGIC TRICK:DIMLY LIT RM, SAY PRODUCE FRO
R	18	7099011-A001/001		CAT POEM:WORD PLACEMENT
R	19	7099012-A001/001	N003501	TOASTER:DRAGON/TOASTER QUALITIES COMPARED
R	20	7099013-A001/001		AD:BEARS NAME SMOKEY
R	21	7099013-A002/002		AD:PURPOSE
R	22	7099013-A003/003		AD:TELLS TO DROWN FIRES
SS	23	7099014-A001/002		BRIAN GREEN APP:NAME
SS	24	7099014-A003/004		BRIAN GREEN APP:BIRTHDATE
SS	25	7099014-A005/006		BRIAN GREEN APP:ADDRESS
SS	26	7099014-A007/008		BRIAN GREEN APP:FATHER
SS	27	7099014-A009/010		BRIAN GREEN APP:TELEPHONE
SS	28	7099014-A011/012		BRIAN GREEN APP:BUS ADD
SS	29	7099014-A013/014		BRIAN GREEN APP:SCHOOL
SS	30	7099014-A015/016		BRIAN GREEN APP:GRADE
SS	31	7099014-A017/018		BRIAN GREEN APP:COUNSELOR
SS	32	7099014-A019/020		BRIAN GREEN APP:SUBJECTS
SS	33	7099014-A021/022		BRIAN GREEN APP:FAILED
SS	34	7099014-A023/024		BRIAN GREEN APP:MISBEHAVE
SS	35	7099016-A001/001		TABLE CONTENTS:MOVIE REV
SS	36	7099016-A002/002		TABLE CONTENTS:SCIENCE
SS	37	7099016-A003/003		TABLE CONTENTS:ARTICLE
SS	38	7099017-A001/002		TV GUIDE:RERUN
SS	39	7099017-A003/003		TV GUIDE:BOTH MOVIE & ZOO
SS	40	7099017-A004/005		TV GUIDE:NO NEW PROG AT 3 ON 4
SS	41	7099017-A006/007		TV GUIDE:TIME OF CARTOONS
SS	42	7099017-A008/008		TV GUIDE:LENGTH OF PROG ON 6
SS	43	7099018-A001/001		INTER VOYAGE: REPT TRAVEL

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
SS	44	7099018-A002/002		INTER VOYAGE: OBERTH DIE
SS	45	7099018-A003/003		INTER VOYAGE: TRANSLATED
SS	46	7099019-A001/001		AUTO:INS MAX AMT MED BILL
SS	47	7099019-A002/002		AUTO:INS MAX AMT INJURED
R	48	7099020-A001/001		TV GUIDE PART:MYSTERY
R	49	7099020-A002/002		TV GUIDE PART:AFTERNOON
R	50	7099020-A003/003		TV GUIDE PART:BOB JOHNSON
R	51	7099021-A001/001		BEAT PARA:PT TO DEFINE
R	52	7099021-A002/002		BEAT PARA:IN COLL ESSAYS
R	53	7099021-A003/003		BEAT PARA:'FINE, NEGLECTED'
R	54	7099021-A004/004		BEAT PARA:ORIGINS OBSCURE
R	55	7099022-A001/001		SUBURBANITES:ABYSS DEBTS
R	56	7099022-A002/002		SUBURBANITES:SECURE DEBTS
R	57	7099023-A001/001		NAYON:GEOG FACTORS
R	58	7099023-A002/002		NAYON:BY 1948 DEPENDENT
R	59	7099023-A003/003		NAYON:WHY SEPARATED
R	60	7099024-A001/001	N013701	OLD MAN:STORY TELLS HOW MAN LOOKS
R	61	7099025-A001/001		BOOK NOT ALL ABOUT PEOPLE
R	62	7099026-A001/001		INC TAX FORM:SINGLE IF DIVORCED 1-79
R	63	7099026-A002/002		INC TAX FORM:NOT FILE JNT
R	64	7099026-A003/003		INC TAX FORM:JNT 1967
R	65	7099027-A001/001		CORP KINDNESS PERSONAL
R	66	7101007-A001/001		COMPOUND WORD CLASSROOM
R	67	7101009-A001/001		MICROSCOPE USED FOR
R	68	7101017-A001/001		PHEASANT MEANS GAME BIRD
R	69	7101019-A001/002		WORDS:MEAN ABATE
R	70	7101019-A003/004		WORDS:MEAN VEHEMENTLY
R	71	7101019-A005/006		WORDS:MEAN INCORRIGIBLE
R	72	7101019-A007/008		WORDS:MEAN MOROSELY
R	73	7101019-A009/010		WORDS:MEAN PROFICIENT
R	74	7101019-A011/012		WORDS:MEAN FURTIVE
R	75	7101019-A013/014		WORDS:MEAN INNUMBERABLE
R	76	7101055-A001/001	N004101	NONSENSE WORD 1:KAG-FIRE
R	77	7101056-A001/001	N014001	NONSENSE WORD 2:TUP-PAPER
R	78	7101057-A001/001		NONSENSE WORD:TRATS SHOES
R	79	7101058-A001/001		NONSENSE WORD:CAGS HANDS
R	80	7101059-A001/001	N009101	NONSENSE WORD 3:HABBIES-DOGS
R	81	7101060-A001/001		NONSENSE WORD:ZUP WATER
R	82	7101061-A001/001		NONSENSE WORD:MARTS
R	83	7101062-A001/001		LUNCH DOOR CAFETERIA
R	84	7101063-A001/001		PRINCIPALS DOOR PICTURE
R	85	7101064-A001/001		SIGN BUS STOP

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	86	7102001-A001/001	N009401	DUAL:WORD BAT-2 MEANINGS FOOLED NELL
R	87	7102004-A001/001	N014101	SENTENCE 1:MOST SENSE-BLEW HOUSE DOWN
R	88	7102005-A001/001		SENTENCE 2:MOST SENSE-GIRL WALKED TO THE
R	89	7102006-A001/001		MOST SENSE BOY WANTED
R	90	7102007-A001/001		DOG ON LEASH HAS SPOTS
R	91	7102008-A001/001	N010301	SNOWMAN:BEST DESCRIPTION-SOMEONE MADE SNO
R	92	7102010-A001/001	N005101	DRAWING:WINNIE SHORTER THAN PAMELA-BEST S
R	93	7102011-A001/001	N010701	SENTENCE 3:MOST SENSE-BALL ROLLED DOWN TH
R	94	7102013-A001/001	N015201	PEOPLE LEARN TO READ: IN SCHOOL
R	95	7102014-A001/001		IM GOING TO THAT MOVIE
R	96	7102015-A001/001		SENTENCE THAT ASKS QUES
R	97	7102029-A001/001	N008701	PICTURE:DOG LYING ON TOP DOGHOUSE-BEST DE
R	98	7102030-A001/001		SIGN FOR BICYCLISTS
R	99	7102031-A001/001		QUIET SIGN HANGING DOOR
R	100	7102032-A001/001		T/F:CHILDREN ARE HORSES
R	101	7102032-A002/002		T/F:CRAYONS OF BRICKS
R	102	7102032-A003/003		T/F:PENCILS FOR WRITING
R	103	7102032-A004/004		T/F:SUN MAKES YOU COLD
R	104	7102032-A005/005		T/F:TOUCH EAR WITH TONGUE
R	105	7102033-A001/001		SIGNS WALKING PERMITTED
R	106	7102034-A001/001		GHOST STORY:MOON IS FLASHLT
R	107	7102035-A001/001	N002702	ATMOSPHERE:SCIENTISTS KNOW MOST ABOUT TRO
R	108	7102036-A001/001		AUTO WRECK:WINGS TURNS
R	109	7102037-A001/001		SHAKESPEARE:DEAF HEAVEN
R	110	7102037-A002/002		SHAKESPEARE:LOVE SAVES
R	111	7102038-A001/001		AUTO WRECK:TERRIBLE CARGO
R	112	7103002-A001/001		WILLY WORM 1:STORY ABOUT A HUNGRY WORM
R	113	7103004-A001/001		EASTER EGGS IN PAST TITLE
R	114	7103012-A001/001		GOTROCKS:WENT MT EVEREST
R	115	7103017-A001/001		SPORTS CAR TURNS CORNERS
R	116	7103019-A001/001		BIRDS:CRY LIKE MOUSE WHEN ANGRY
R	117	7103020-A001/001	N003901	SELFISH PERSON:DESCRIPTION IN PASSAGE
R	118	7103021-A001/001	N003401	YOUNG GARDENERS:IN CENTRAL PARK-BEST
R	119	7103025-A001/001	N001101	PICTURE:CEREAL WITH TOY INSIDE IS PAX
R	120	7103026-A001/001		WHICH BUBBLE GUM SWEET
R	121	7103027-A001/001		ZOO SIGN DANGEROUS ANIMAL
R	122	7103028-A001/001		SIGN FOR PEDESTRIANS
R	123	7103029-A001/001		CAT POEM:BUTTONS SCATTER
R	124	7103030-A001/001		ENG MUFFINS:BAKING TIME
R	125	7103031-A001/001		SCARLET FEVER:HOW FEEL
R	126	7103032-A001/001		MT EVEREST:2 HEIGHTS
R	127	7103033-A001/001		COLORADO MOUNTAINS PASS

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	128	7103034-A001/001		WIND BOAT STORMY DAY SEA
R	129	7103035-A001/001		H KELLER:ACCOMPLISHMENTS
R	130	7103036-A001/001		GREG GOTROCKS:SENT TO NEPAL TO CHECK ON O
R	131	7103037-A001/001		PERSON LIKES SPY STORIES
R	132	7103038-A001/001		INCOME TAX:MAX FOR SEPARATE IS $500. EACH
R	133	7103039-A001/001		MENTAL RETARD:AD PURPOSE-ENCOURAGE HIRING
R	134	7103041-A001/001		KOLA COUPON:APPEALS TO EVERYONE
R	135	7103041-A002/002	N001301	KOLA COUPON:GOOD FOR ANY SIZE CARTON
R	136	7103041-A003/003	N001302	KOLA COUPON:USE ON NOV. 10, 1970
R	137	7103041-A004/004	N001303	KOLA COUPON:PAYMENT IS 12 CENTS
R	138	7103042-A001/001		9 OUT OF 10 AMERS DEBT
R	139	7103042-A002/002		INCOME INCREASED 50%
R	140	7103043-A001/001		ENG MUFFINS:BAKED GRIDDLE
R	141	7103044-A001/001	N001701	BOOK CLUB:SHIPPING COSTS HIGHER IN CANADA
R	142	7103044-A002/002	N001702	BOOK CLUB:SEND NO MONEY TILL BILLED
R	143	7103044-A003/003	N001703	BOOK CLUB:BUY 6 MORE
R	144	7103045-A001/001	N005201	TRAFFIC:APPEAR IN COURT TO PLEAD NOT GUIL
R	145	7103045-A002/002	N005202	TRAFFIC:FINE-$3.00
R	146	7103045-A003/003	N005203	TRAFFIC:PAY FINE BY THURS, JUNE 11
R	147	7103046-A001/001		FILM NOTICE:DAMAGE REPL
R	148	7103046-A002/002		FILM NOTICE:COLOR CHANGES
R	149	7103047-A001/001		H KELLER:WHEN LOST SIGHT
R	150	7103048-A001/001		SILKY 1:PLAYED INSTRUMENTS
R	151	7103049-A001/001		FARMER BROWN:FARMERS KNOW
R	152	7103049-A002/002		FARMER BROWN:WRITER'S IDEA
R	153	7103050-A001/001		MARTIAN POLAR CAPS:DISCOVERED MORE THAN 2
R	154	7103051-A001/001		BUG SPRAY:NOT KILL FLIES
R	155	7103051-A002/002		BUG SPRAY:HOLD 10 INCHES
R	156	7103052-A001/001		AUTO WRECK:PEOPLE DEAD
R	157	7103053-A001/001		SILKY:DIDN'T LIKE RAIN
R	158	7103054-A001/001		WIND BOAT WEATHER WAS WET
R	159	7103055-A001/001		SKIING:NO ACCOMMADATIONS
R	160	7103056-A001/001		WIND BOAT # OF PEOPLE 2
R	161	7103057-A001/001		CAT/BIRD COMIC:POINT
R	162	7103058-A001/001		H KELLER:EXTENT LECT TOURS
R	163	7103059-A001/001		FRANGIBLES:COMMUNICATE
R	164	7103060-A001/001		SCARLET FEVER:OTHER INFEC
R	165	7103061-A001/001		HOW SPORTS CARS DIFFER
R	166	7103062-A001/001		POISON IVY:WASH TO AVOID
R	167	7103062-A002/002		CALAMINE LOTION SOOTHES
R	168	7103062-A003/003		BORIC ACID FOR EYELIDS
R	169	7127001-A001/001	N003801	SCOTT:BEST TITLE-SCOTT'S PLAN

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	170	7127001-A002/002	N003802	SCOTT:6 WEEKS BETWEEN DEPOTS
R	171	7127001-A003/003	N003803	SCOTT:CACHE-PLACE FOR STORING THINGS
R	172	7127002-A001/001		SLEEKY:HOW MANY OTTERS
R	173	7127002-A002/002		SLEEKY:WORD CHATTER MEAN
R	174	7127003-A001/001	N002101	VIRUSES:DIFFICULT TO STUDY
R	175	7127003-A002/002	N002102	VIRUSES:CLOTHE IDEA-GIVE PROOF TO SUPPORT
R	176	7127004-A001/001		SUBURBANITES:SELF ENTRAP
R	177	7127004-A002/002		SUBURBANITES:BUDGETISM
R	178	7127005-A001/001		FARMER BROWN:MAIN IDEA
R	179	7127005-A002/002		FARMER BROWN:CHANGE ENVIR
R	180	7127006-A001/001		PERSIAN GULF OYSTERS
R	181	7127007-A001/001		SOCIAL SCI:WRONG TO NEGLECT BEHAV & CULT
R	182	7127009-A001/002		TAA HOSTESS:COMPANY
R	183	7127009-A003/004		TAA HOSTESS:JOB
R	184	7127009-A005/005		TAA HOSTESS:QUALIFICATION 1
R	185	7127009-A006/006		TAA HOSTESS:QUALIFICATION 2
R	186	7127009-A007/007		TAA HOSTESS:TOP SALARY
R	187	7127009-A008/008		TAA HOSTESS:HOW APPLY
R	188	7127009-A009/009		TAA HOSTESS:EOE
R	189	7201002-A001/001		AMOS ANT:WENT TO PARK FIRST
R	190	7201003-A001/001		WIND BOAT PUSH FIRST WENT SEA
R	191	7201013-A002/003		SEQUENCE CARTOONS
R	192	7201014-A001/001		SEA FEVER:POET ASKS FOR
R	193	7201023-A001/001		H KELLER:WHEN STUDY PROBS
R	194	7201024-A001/001		ENG MUFFINS:4 INGREDIENTS
R	195	7201025-A001/001		SCARLET FEVER:OTHER DISEA
R	196	7202003-A001/001	N002701	ATMOSPHERE:4 WORDS CUE-FIRST,NEXT,ABOVE,F
R	197	7202008-A001/001		EVENTS:BEFORE MEETING WENT TO CONF ROOM
R	198	7203002-A001/001		GHOST STORY:MOOD FRIGHT
R	199	7203003-A001/001		GHOST STORY:ADD MYSTERY
R	200	7203006-A001/001		H KELLER:IN CHRONOLOGICAL
R	201	7203009-A001/001		ANGRY:TONE BEST DESCRIBED AS ANGRY
R	202	7203009AA001/001		CANNOT TOLERATE ANGRY
R	203	7203010-A001/001		TURTLE POEM:UNUSU PT OF VIEW
R	204	7203011-A001/001		FLIES EXAGGERATING SIZE
R	205	7203012-A001/001		SENTENCE TONE SATIRICAL
R	206	7203013-A001/001		TURTLE POEM:QUICK CLAMPS
R	207	7203043-A001/001		GHOST STORY:WIND SOUNDS
R	208	7203044-A001/001		FLIES:AUTHOR WANTS YOU TO THINK IT'S FUNN
R	209	7203045-A001/001		FISH WALK TO MAKE LAUGH
R	210	7203046-A001/001		SKIING:LOVE OF
SS	211	7203047-A001/001	N011701	WHICH WORD COMES FIRST IN DICTIONARY- FLE

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	212	7203048-A001/001		TURTLE POEM:CONTENTED
R	213	7203049-A001/001		STATIC CULTURE EARNEST
R	214	7203050-A001/001		FISH WALK FUNNY STORY
R	215	7203051-A001/001		WIND WHISTLED SOUND PAIR
R	216	7203052-A001/001		CAT/BIRD COMIC:TONE
R	217	7203053-A001/001		SPEAKER ATTITUDE EXASP
R	218	7203054-A001/001		GARLIC SENTENCES 5 AND 6
R	219	7227001-A001/001		FRANGIBLES:MAIN PURPOSE
R	220	7301002-A001/001		TIMOTHY:TIME OF YEAR
R	221	7301004-A001/001	N009601	TIMOTHY 1:SITTING ON STEPS
R	222	7301006-A001/001		TIMOTHY:MEN WASHING CARS
R	223	7301007-A001/001		TIMOTHY:GIRLS JUMP ROPE
R	224	7301011-A001/001	N014201	TIMOTHY 2:TEENAGERS STANDING IN CIRCLES
R	225	7301012-A001/001	N012901	TIMOTHY 3:TEENAGERS TALKING ABOUT HEAT
R	226	7301014-A001/001		TIMOTHY:WORKMEN TEARING
R	227	7301019-A001/002		J DOUGLAS:3 WOMEN IN ROOM
R	228	7301019-A003/004		J DOUGLAS:RUNNING AWAY
R	229	7301020-A001/002		LONE DOG:5 WORDS(1)
R	230	7301020-A003/004		LONE DOG:5 WORDS(2)
R	231	7301020-A005/006		LONE DOG:5 WORDS(3)
R	232	7301020-A007/008		LONE DOG:5 WORDS(4)
R	233	7301020-A009/010		LONE DOG:5 WORDS(5)
R	234	7301020-A011/012		LONE DOG:2 THINGS DOES(1)
R	235	7301020-A013/014		LONE DOG:2 THINGS DOES(2)
R	236	7301022-A001/002		ZEKE:PLACES LIVED 2
R	237	7301022-A003/004		ZEKE:LIVES NOW HARLEM
R	238	7301022-A005/006		ZEKE:HOUSE BROWNSTONE
R	239	7301022-A009/010		ZEKE:TOPMOST FLOOR
R	240	7301027-A001/002		J DOUGLAS:CITY BROOKLYN
R	241	7301027-A003/004		J DOUGLAS:MONTH NOVEMBER
R	242	7301027-A005/006		J DOUGLAS:DAY MONDAY
R	243	7301071-A001/001	N014501	CONNECT DOTS:ALONG LINE, CONNECT DOTS
R	244	7301071-A002/002	N014502	CONNECT DOTS:DRAW LINE TO TOUCH CIRCLES
R	245	7301071-A003/003	N014503	CONNECT DOTS:WRITE 3 IN EACH CIRCLE
R	246	7302001-A001/001		OVAL:FILL IN OVAL BELOW LETTER E
R	247	7302002-A001/002		CONNECT DOTS SOLID LINE A
R	248	7302002-A003/004		WRITE WORD CAT ON LINE
R	249	7302002-A005/006		LINE TO CONNECT 2 AND 7
R	250	7302002-A007/008		CONNECT DOTS SOLID LINE D
R	251	7302004-A001/001		EVER VISITED MOON
R	252	7302004-A002/002		NEVER VISITED MOON
R	253	7302005-A001/001		FIGURE MADE WITH 3 LINES

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	254	7302008-A001/002		WRITE ZERO NOT YOUR AGE
R	255	7302009-A001/002		DRAW HORIZONTAL LINE
R	256	7302009-A003/004		DRAW 2 CIRCLES
R	257	7302009-A005/006		DRAW ANOTHER CIRCLE ABOVE
R	258	7302009-A007/008		CONNECT CENTERS OF 3
R	259	7302009-A009/010		DRAW VERTICAL LINE
R	260	7302012-A001/002		SHAPES:3 IN LARGE CIRCLE
R	261	7302012-A003/004		SHAPES:2 IN SMALL SQUARE
R	262	7302012-A005/006		SHAPES:7 IN LARGE TRIAN
R	263	7302012-A007/008		SHAPES:4 IN SMALL CIRCLE
R	264	7302012-A009/010		SHAPES:5 IN LARGE SQUARE
SS	265	7302213-A001/001	N012101	CODE:WHAT DOES HPPE ACTUALLY SPELL-GOOD
SS	266	7303002-A001/001	N006501	FIND GUIDE:OPTIONAL BETWEEN OPPRESS-ORACL
SS	267	7303003-A001/001	N012601	FIND:BEST PLACE FIND ROTOR-DICTIONARY
R	268	7303004-A001/001		DOG FOOD LABELS PROTEIN
SS	269	7303005-A001/001		ENCYCLOPEDIAS 1:EGGS IN VOL 3
SS	270	7303006-A001/001		ESKIMOS LOOK IN INDEX
SS	271	7303007-A001/001	N011801	ENCYCLOPEDIAS 2:WASHINGTON IN VOL 11
SS	272	7303008-A001/001		ENCYCLOPEDIA:WINDMILLS
SS	273	7303010-A001/001		MAP:NORTHTOWN CLOSER TO HOPE
SS	274	7303010-A002/002		MAP:CAN DRIVE TO FALLS CITY
SS	275	7303010-A003/003		MAP:HOPE CLOSEST TO CENTERVL
SS	276	7303010-A004/004		MAP:CENTERVILLE FARTHER WEST
SS	277	7303010-A005/005		MAP:HWY 20 S OF RIVER
SS	278	7303012-A001/001	N012001	FIND:DECLARATION OF INDEPENDENCE IN ENCYC
R	279	7303013-A001/001	N012301	PIC:3 PARTS MUSHROOM-CAP,STEM,GILLS
SS	280	7303014-A001/001		JONES PHONE NUMBER
R	281	7303017-A001/001	N004501	AREA CODE:INFO NY-1-212-555-1212
R	282	7303017-A002/002	N004502	AREA CODE:SYRACUSE-1-315-255-6011
SS	283	7303018-A001/002	N006901	NEWS:TV SCHEDULE-PG 22
SS	284	7303018-A003/004	N006902	NEWS:WEATHER FORECAST-PG 12
SS	285	7303018-A005/006	N006903	NEWS:STOCK AVERAGES-PGS 29-31
SS	286	7303018-A007/008		NEWS:QUESTION **NOT SCORED**
SS	287	7303018-A009/009		NEWS:BRIDGE INFO GIVEN
R	288	7303019-A001/001	N001201	LONG DIST:RATE ON CALL-LOWER EVENING RATE
R	289	7303019-A002/002	N001202	LONG DIST:PERSON CALLS DIFF-OPR ASSISTED
SS	290	7303023-A001/001		FATAL ACCIDENTS 2 TO 3AM
R	291	7303026-A001/001		HELP WANTED:AD HOURS AM
R	292	7303026-A002/002		HELP WANTED:AD HOURS PM
R	293	7303026-A003/004		HELP WANTED:AD AGE REQ
R	294	7303026-A005/005		HELP WANTED:AD SALARY
SS	295	7303027-A001/001	N006101	WIND SYMBOLS:FOR 35 KNOTS-SYMBOL 3

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
SS	296	7303030-A001/001		REPORT CARD:PERIOD
SS	297	7303031-A001/001		REPORT CARD:IMPROVE SCI
SS	298	7303033-A001/001		REPORT CARD:ALGEBRA PROB
SS	299	7303034-A001/002	N005901	CARDCAT:CALL NUMBER-WRITE-IN ANSWER
SS	300	7303034-A003/003	N005902	CARDCAT:PICTURES INDIC BY "ILLUS"
SS	301	7303035-A001/001	N006801	MAP:SPANISH IN SOUTH
SS	302	7303035-A002/002	N006802	MAP:GRP IN ALASKA-NOT ENOUGH INFO
SS	303	7303037-A001/001	N006301	CLOTHES SIZES:SHOE SIZE 8-40-1
SS	304	7303037-A002/002	N006302	CLOTHES SIZES:38 SWEATER-44
SS	305	7303042-A001/001	N006701	SCI INDEX:WOLVES FIRST IN BOOK
SS	306	7303043-A001/001		REPORT CARD:FOREIGN LANG
SS	307	7303044-A001/001		ACC CHART:INCONCLUSIVE
SS	308	7303045-A001/001		SCI INDEX:EARTHWORMS INFO
SS	309	7303050-A001/001	N005801	ENGLISH DIC:BOOK TELLS WORD MEANINGS-DICT
R	310	7303051-A001/001	N002201	PHONE BILL:FEB 14 CALL FROM ATHENS, GA
R	311	7303051-A002/002	N002202	PHONE BILL:FEB 14 CALL TO ST PAUL, MN
R	312	7303051-A003/003	N002203	PHONE BILL:FEB 14 CALL COST $.75
R	313	7303052-A001/001		CLOCK BIG HAND BET 12 & 1
R	314	7303054-A001/001		FISHING:METHOD NOT PERMITTED-USE MORE THA
R	315	7303054-A002/002		FISHING:FOR MULLET-CAN USE ALL METHODS
R	316	7303055-A002/002		NUCLEAR BURSTS:IMM DANGER
R	317	7303055-A003/003		NUCLEAR BURSTS:SKIN BURNS
R	318	7303056-A001/002		WIN-EM-ALL:DEALER CHOSEN
R	319	7303056-A003/004		WIN-EM-ALL:ADULTS & CHILD
R	320	7303056-A005/006		WIN-EM-ALL:NO MORE CARDS
R	321	7303056-A007/008		WIN-EM-ALL:1ST PLAYER
R	322	7303057-A001/002		WIN-EM-ALL:DEALS FIRST
R	323	7303057-A003/004		WIN-EM-ALL:HOW MANY PLAY
R	324	7303057-A005/006		WIN-EM-ALL:TIE TRICK
R	325	7303057-A007/008		WIN-EM-ALL:WINNER
SS	326	7303058-A001/001		ACCIDENT STATE BEY FACTS
SS	327	7303059-A001/001		SCHEDULE OTHER CHOICES GO
SS	328	7327001-A001/001		ST PAUL:CALL LAKEVILLE CHARGE
SS	329	7327001-A002/002		ST PAUL:CALL MAPLE PLAIN CHARGE
SS	330	7327001-A003/003		ST PAUL:CALL MINNEAPOLIS NO CHRG
SS	331	7327001-A004/004		ST PAUL:CALL NORTH AREA NO CHRG
SS	332	7327001-A005/005		ST PAUL:CALL SHAKOPEE CHARGE
SS	333	7327001-A006/006		ST PAUL:CALL SOUTHWEST AREA NO CHRG
SS	334	7327001-A007/007		ST PAUL:CALL WHITE BEAR LAKE NO CHRG
SS	335	7327001-A008/009		ST PAUL:CALL WHAT AREA 533-0221-NOTHWST(H
R	336	7401001-A001/001		SILKY SPIDER 2:SILKY WAS HUGE-BEST DESCRI
R	337	7401003-A001/001		TURTLE POEM:SPEAKING

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	338	7401005-A001/001		SILKYS WEB VERY BIG
R	339	7401007-A001/001		H KELLER:SULLIVAN METHOD
R	340	7401010-A001/001		SILKY:LIKES BEAN SOUP
R	341	7401011-A001/001		WILLY 2:ATE APPLE
R	342	7401016-A001/001	N004801	SILKY 3:WISHED MORE HAIR
R	343	7401019-A001/001		SILKY:LIKED CUDDLES
R	344	7401022-A001/001		SILKY:FLIES AS PLAYMATES
R	345	7401024-A001/001		BEST TITLE:A TASTE FOR READING
R	346	7401030-A001/001		FRANGIBLES:SEEMS FALSE
R	347	7401032-A001/001		HORSEPOWER WITHOUT SENSE
R	348	7401066-A001/001		FISH PICTURE ABOUT TO EAT
R	349	7401067-A001/001	N003001	SUPR COURT:CONSTITUTION DESCRIPTION-BRIEF
R	350	7401067-A002/002	N003002	SUPR COURT:DIFFICULT RESPON FOR COURT MEM
R	351	7401067-A003/003	N003003	SUPR COURT:"THEIR" REFERS TO PROVISIONS
R	352	7401068-A001/001	N008801	YVONNE'S DOLL:COULDN'T FIND-UNDER PORCH
R	353	7401069-A001/001		FINISH 2ND STORY LIKE 1ST
R	354	7401070-A001/001		CAT/BIRD COMIC:BIRD WOULD SAY
R	355	7401071-A001/001	N010201	DESCRIPTION 1:CLOWN DESCRIBED IN PASSAGE
R	356	7401072-A001/001	N013301	DESCRIPTION 2:UNHAPPY PERSON DESCRIBED IN
R	357	7401073-A001/001	N009901	DESCRIPTION 3:PERSON HAS SEEN TOY MANY TI
R	358	7401074-A001/001	N001401	VERSE:DECK OF CARDS DESCRIBED IN POEM
R	359	7401075-A001/001		VERSE:CLOCK
R	360	7401076-A001/001		VERSE:FLAG
R	361	7401077-A001/001		VERSE:EYEGLASSES
R	362	7401078-A001/001		TOMMY AND SAMMY FIGHT
R	363	7401079-A001/001		FRANGIBLES:ENTER OBJECT
R	364	7401080-A001/001		STATIC CULTURE ATTITUDES
R	365	7401081-A001/001		WIND BOAT HELP NOW RESCUE
R	366	7401082-A001/001		CHRISTMAS NEAR COATS
R	367	7401083-A001/001		POEM:UNSURE ATTITUDE
R	368	7401084-A001/001	N011201	DOGS' QUAL:BITTEN BY DOG, DISAGREE
R	369	7401085-A001/001		CHRISTMAS SHOPPING LAST
R	370	7401086-A001/001		CHRISTMAS STORY DEC 21
R	371	7402020-A001/001	N002401	MOSQUITO:SIZE MOSQUITOES EXAGGERATED
R	372	7402021-A001/001	N009201	PUZZLE 1:BIRD DESCRIBED IN PUZZLE
R	373	7402022-A001/001		PUZZLE 2:WORM DESCRIBED IN PUZZLE
R	374	7402023-A001/001	N009801	PUZZLE 3:CHAIR DESCRIBED IN PUZZLE
R	375	7403007-A001/001	N004901	COLORADO:GOLD DISCOVERY DOESN'T BELONG
R	376	7403018-A001/001		BERT & ART NOT BOTH RIGHT
R	377	7403019-A001/001	N002501	MARY:WILL GET MONEY FROM NEITHER
R	378	7502012-A001/001		AMOS ANT:MAKE-BELIEVE
R	379	7503001-A001/001	N011101	KIND OF BK:ATMOSPHERE FROM SCIENCE BOOK

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	380	7503004-A001/001		PASSAGE WITH AGE CONFLICT
R	381	7503005-A001/001		SKIING:PERSONAL POINT
SS	382	7503009-A001/002		BANK CHECK:WHO RECEIVES
SS	383	7503009-A003/004		BANK CHECK:NUMBER
SS	384	7503009-A005/005		BANK CHECK:CANT BE CASHED
R	385	7503044-A001/001		ATMOSPHERE:OPINION
R	386	7503045-A001/001	N003301	BOBBY:SAYS TALL IS SMART
R	387	H201000-A001/001	N008601	CRICKETS: MAKE SOUNDS BY RUBBING WINGS
R	388	H201000-A002/002	N008602	CRICKETS: WHICH MAKE CHIRPING SOUNDS-ONLY
R	389	H201000-A003/003	N008603	CRICKETS: WHERE ARE EARS - IN FRONT LEGS
R	390	H202000-A001/001		EXTINCT:PAY A PRICE MEANS GIVE UP IN RETU
R	391	H202000-A002/002		EXTINCT:MANY MARSUPIALS IN AUSTRALIA-NO C
R	392	H202000-A003/003		EXTINCT:WELL ADAPTED SPECIES MAYBE OVERSP
R	393	H202000-A004/004		EXTINCT:MARSUPIALS & PLACENTALS-MANY DIFF
R	394	H204000-A002/002	N010902	STARS UNSEEN:STAR BECOMES DEAD BY USING U
R	395	H204000-A003/003	N010903	STARS UNSEEN:MAIN IDEA-STARS EXIST-WE CAN
R	396	H204000-A004/004	N010904	STARS UNSEEN:DEAD STARS BIG & HEAVY-PUSH
R	397	H204000-A005/005		STARS UNSEEN:RADIO STAR-AREA FILLED ELEC.
R	398	H205000-A001/001	N010501	QUICKSAND:HOW TEST FOR IT-POKE WITH A STI
R	399	H205000-A002/002	N010502	QUICKSAND:MAIN PURPOSE-TO TELL WAYS AVOID
R	400	H205000-A003/003	N010503	QUICKSAND:IT IS SOUPY SAND YOU CAN'T STAN
R	401	H205000-A004/004	N010504	QUICKSAND:IF STEP IN,LIE ON BACK & STRETC
R	402	H206000-A001/001	N011301	SKUNK CABBAGE:NAME-SMELLS LIKE SKUNK,LOOK
R	403	H206000-A002/002	N011302	SKUNK CABBAGE:HARD TO SEE-HIDDEN UNDER HO
R	404	H222000-A001/001	N001601	1ST AM:BITTER WINTER-EXTREMELY COLD
R	405	H222000-A002/002	N001602	1ST AM:ICE AGE PEOPLE DEPENDED ON ANIMALS
R	406	H222000-A003/003		1ST AM:KIND OF PEOPLE-WANDERERS NEEDING A
R	407	H222000-A004/004	N001603	1ST AM:NO LAND BRIDGE NOW-COVERED WITH WA
R	408	H222000-A005/005	N001604	1ST AM:MAIN PURPOSE-EXPLN ICE AGE SETTLER
R	409	H224000-A002/002		FORD:1ST CARS COSTLY BECAUSE TOOK TIME TO
R	410	H224000-A003/003		FORD:PROFIT-MONEY AFTER EXPENSES PAID
R	411	H224000-A004/004		FORD:WORK MADE EASIER BY RAISING ASSEMBLY
R	412	H225000-A001/001		RUSS PORTS:WHY FEW USABLE-WATER FROZEN MO
R	413	H225000-A002/002		RUSS PORTS:BALTIC ATTRACTIVE-LINK INTERIO
R	414	H225000-A003/003		RUSS PORTS:GREAT NO.WAR & JAPAN WAR TO WI
R	415	H225000-A004/004		RUSS PORTS:MAIN PURPOSE-DISCUSS EFFORTS G
R	416	H225000-A005/005		RUSS PORTS:AVENUE IN SENTENCE MEANS ROUTE
R	417	H241000-A001/001	N004401	NAOMI JAMES:HOW LONG ON SAILING TRIP- 272
R	418	H241000-A002/002	N004402	NAOMI JAMES:IMPORTANCE OF TRIP-BROKE WORL
R	419	H241000-A003/003	N004403	NAOMI JAMES:WORST PART OF TRIP- BAD STORM
R	420	H243000-A002/002		COUSTEAU:PEOPLE SEEK ADVENTURE TO FIND OU
R	421	H243000-A003/003		COUSTEAU:LOWER ODDS-REDUCE RISKS

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	422	H243000-A004/004		COUSTEAU:CALM AFTER SHARK-WANT CLEAR PICT
R	423	H243000-A005/005		COUSTEAU:AQUALUNG IMPORTANT-ALLOWS FREER
R	424	H262000-A002/002	N015502	CHAMONIX:WHY SO LONG TO REACH-WINDS TOO S
R	425	H262000-A003/003	N015503	CHAMONIX:DEVOUASSOU-MAN WHO FOUND CLIMBER
R	426	H262000-A004/004	N015504	CHAMONIX:DESMAISON SURVIVE BY MENTAL/PHYS
R	427	H262000-A005/005	N015505	CHAMONIX:WHY DESMAISON CRY-OVERCOME SUFFE
R	428	H263000-A002/002	N002902	SOCCER:MOST POPULAR BECAUSE PLAYED BY MIL
R	429	H263000-A003/003	N002903	SOCCER:KING ED WANTED TO OUTLAW-PRACTICE
R	430	H263000-A004/004	N002904	SOCCER:CALLED FOREIGN-IMMIGRANTS PLAYED I
R	431	H263000-A005/005	N002905	SOCCER:INTRO TO ENGLISH BY ROMANS
R	432	H263000-A006/006	N002906	SOCCER:PELE MASTER-FOOLED OPPONENTS BY FA
R	433	H265000-A001/001	N002001	WISH COULD FLY:GOSSAMER CONDOR 1ST MUSCLE
R	434	H265000-A002/002	N002002	WISH COULD FLY:BIKE RACER, BRYAN ALLEN FL
R	435	H265000-A003/003	N002003	WISH FLY:MACCREADY PLANE DIFF-SIMPLER,LIG
R	436	H266000-A001/001		NO NICE BEAR:IN PAST SMOKEY OFFERED POLIT
R	437	H266000-A002/002		NO NICE BEAR:CHNGD IMAGE BECAUSE MORE FOR
R	438	H268000-A001/001	N013201	BULLFIGHT:BULL CHARGES CAPE MOTION
R	439	H269000-A002/002	N010102	SANDWICH:NAMED AFTER PERSON WHO INVENTED
R	440	H269000-A003/003	N010103	SANDWICH:WANTED MEAT IN BREAD TO EAT AND
R	441	H282000-A001/001		LABELS:ASPIRIN/5-YR-OLD,TAKE 1/2 TABLET
R	442	H282000-A002/002		LABELS:EXTERNAL USE-DO NOT DRINK
R	443	H282000-A003/003		LABELS:ANTIDOTE/ANTIDOTE-A TREATMENT
R	444	H284000-A001/001	N003201	SUMMER JOB:SOC SECURITY APPLIC AT BANK OR
R	445	H284000-A002/002	N003202	SUMMER JOB:BEST TIME TO FIND JOB-BEFORE M
R	446	H284000-A003/003	N003203	SUMMER JOB:NEED SS CARD TO GET INTERVIEW
R	447	H284000-A004/004	N003204	SUMMER JOB:REFERENCES-PEOPLE WHO KNOW APP
R	448	H286000-A001/001	N004701	CARRIER AD:IF INTEREST & MEET REQRMNTS-CA
R	449	H286000-A002/002	N004702	CARRIER AD:8 YR OLDS TOO YOUNG FOR JOB
R	450	H286000-A003/003	N004703	CARRIER AD:MUST DELIVER PAPERS BY 7 EACH
R	451	H287000-A001/001	N013501	CRIME:HARD TO PROVE OWN BIKE IF REPAINTED
R	452	H287000-A002/002	N013502	CRIME:MAIN PURPOSE-TO GIVE SECRET WAY TO
R	453	H301000-A001/001		OLYMPIC AD:MAIN PURPOSE-ENCOURAGE SUPPORT
R	454	H301000-A002/002		OLYMPIC AD:SENTENCE MEANS CITIZENS PROVID
R	455	H301000-A003/003		OLYMPIC AD:OLYMPIC GOLD-MEDALS FOR WINNER
R	456	H302000-A001/001		INTELL:PURPOSE-BRAIN SIZE NOT DETERMINE I
R	457	H302000-A002/002		INTELL:USES FACTS & FIGURES
R	458	H302000-A004/004		INTELL:REFERS TO EXPERTS OF SAME OPINION
R	459	H302000-A005/005		INTELL:PRETENDS AGREE WITH OTHER POINT OF
R	460	H304000-A002/002		FITNESS:PHYS FITNESS HELPS CHILD BE HEALT
R	461	H304000-A003/003		FITNESS:AD PURPOSE-CONVINCE PARENTS KIDS
R	462	H304000-A004/004		FITNESS:PARENTS GET INFO FROM SCHOOL OR R
R	463	H403000-A001/001	N007301	BRIDGER:KIND OF PEOPLE WERE MTN MEN-FUR T

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	464	H403000-A002/002	N007302	BRIDGER:BEST DESCRIBE STORIES-STRETCHED T
R	465	H403000-A003/003	N007303	BRIDGER:SIMILE-PONDS OF MUD BOILING LIKE
R	466	H403000-A004/004	N007304	BRIDGER:WHO DISCOVERED LAND NOW YELLOWSTO
R	467	H403000-A005/005	N007305	BRIDGER:SHOT MISSED ELK BECAUSE ELK OUT O
R	468	H403000-A006/006	N007306	BRIDGER: HYPERBOLE- LAKES THAT HAD NO BOT
R	469	H404000-A001/001	N013101	THE COLD:BOY LEFT SHADOW-FROZE TO SIDE OF
R	470	H404000-A002/002	N013102	THE COLD:GIRLS FIGHT WITH MELTED WORDS
R	471	H404000-A003/003	N013103	THE COLD:DUCKS FLY AWAY WITH POND
R	472	H404000-A004/004	N013104	THE COLD:WRITER MAKES STORY SOUND PLAYFUL
R	473	H405000-A001/001	N001501	NUTS:DEVIL PUT PEARL IN WALNUT
R	474	H405000-A002/002	N001502	NUTS:FARM WIFE WAS CLEVER AND PRACTICAL
R	475	H405000-A003/003	N001503	NUTS:WANTED TRICK SOMEONE INTO CRACKING W
R	476	H405000-A004/004	N001504	NUTS:PLAN WRONG-WOMAN DIDN'T CRACK WALNUT
R	477	H406000-A001/001		GOOD DOG:DOG'S DEATH DESCRIBED-PAINLESS &
R	478	H406000-A002/002		GOOD DOG:SIMILE-LOOKED WITH EYES LIKE BUL
R	479	H406000-A003/003		GOOD DOG:MAN WHO LIVED WITH DOG-CARING &
R	480	H406000-A005/005		GOOD DOG:HYPERBOLE-COULD EAT 100 LOAVES O
R	481	H408000-A002/002		BOKUDEN:SUGGESTED HE & SWORDSMAN FIGHT ON
R	482	H408000-A003/003		BOKUDEN:THEME-CLEVERNESS OVERCOME PHYSICL
R	483	H408000-A004/004		BOKUDEN:BRAGGART MEANS BOASTFUL
R	484	H408000-A005/005		BOKUDEN:AFTER REMOVED JACKET-SHOVED BOAT
R	485	H408000-A006/006		BOKUDEN:MUTEKATSU HIGHEST SKILL BECAUSE N
R	486	H409000-A002/002		OLD MAN:GRANDFATHER EATS SLOPPY-OLD & WEA
R	487	H409000-A003/003		OLD MAN:GRANDFATHER FELT SAD WHEN NOT AT
R	488	H409000-A004/004		OLD MAN:DISGUSTED MEANS ANNOYED
R	489	H409000-A005/005		OLD MAN:MAN & WIFE CRY BECAUSE WAY TREAT
R	490	H409000-A006/006		OLD MAN:GRANDSON MAKE TROUGH GIVE PARENTS
R	491	H412000-A002/002		FUN:MARGIE LEARNED ENGLISH AT HOME BY TV
R	492	H412000-A003/003		FUN:BOOK WAS ABOUT SCHOOL
R	493	H412000-A004/004		FUN:STORY TAKES PLACE IN FUTURE
R	494	H413000-A001/001	N008201	COW-TAIL:OGALOUSSA WAS KILLED WHILE HUNTI
R	495	H413000-A002/002	N008202	COW-TAIL:THEME-PERSON NOT DEAD TILL FORGO
R	496	H413000-A003/003	N008203	COW-TAIL:OGALOUSSA IS WISE,FAIR FATHER
R	497	H413000-A004/004	N008204	COW-TAIL:OGALOUSSA SHAVED HEAD-RETURNED F
R	498	H413000-A005/005	N008205	COW-TAIL:PULI GOT SWITCH-ASKED ABT FATHER
R	499	H416000-A002/002	N010002	DOG & SHADOW:SAW HIMSELF IN THE STREAM
R	500	H416000-A003/003	N010003	DOG & SHADOW:TEACHES LESSON-GREED DOESN'T
R	501	H417000-A001/001		FLOWERS:WILLIE THOUGHT OK STEAL-DEAD MAN
R	502	H417000-A002/002		FLOWERS:IF GLORIA KNEW-WOULDN'T LIKE THEM
R	503	H417000-A003/003		FLOWERS:AT END,MAN FROM GRAVE CAME TO SEE
R	504	H417000-A004/004		FLOWERS:WILLIE SAID BOUGHT-THOUGHT MOM TA
R	505	H418000-A001/001		HUMBUG:STRIKING IT RICH-FINDING LOTS OF G

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
R	506	H418000-A002/002		HUMBUG:GOLD WAS SUPPOSEDLY LYING ON THE G
R	507	H418000-A003/003		HUMBUG:GOT NAMED BECAUSE PEOPLE WERE FOOL
R	508	H419000-A001/001	N010601	THAD:CANDIDATES FOR PRES NOT ALLOWED GIVE
R	509	H419000-A002/002	N010602	THAD:MAGGIE THOUGHT THAD GOOD BUT NEED HE
R	510	H419000-A003/003	N010603	THAD:MASSIVE STAMPEDE-LOT OF PEOPLE RUSHI
R	511	H419000-A004/004	N010604	THAD:EXAGGERATED-CAN DO EVERYTHING IN YEL
R	512	H419000-A005/005	N010605	THAD:MAGGIE FIRST HELPED THAD WITH SPEECH
R	513	H422000-A001/001	N013401	FROM THE PLANET:BOTCHIK FELT ANNOYED AND
R	514	H422000-A002/002	N013402	FROM THE PLANET:THOUGHT NO LIFE-THICK CLO
R	515	H422000-A003/003	N013403	FROM THE PLANET:IN GLASS CAGE WAS A HUMAN
R	516	H441000-A001/001		TOMATO MAN:WRITERS BROUGHT CAMERA TO HIS
R	517	H441000-A002/002		TOMATO MAN:PUTTIN' UP 'MATERS MEANS CAN T
R	518	H441000-A003/003		TOMATO MAN:AFTER PEELING,SLICE & CORE TOM
R	519	H442000-A001/001	N008101	CLOSING:PUN-DOORMAN AT PLAZA? NO
R	520	H442000-A002/002	N008102	CLOSING:PUN-MORE THAN 50 YEARS? NO
R	521	H442000-A003/003	N008103	CLOSING:PUN-END SWINGING CAREER? YES
R	522	H442000-A004/004	N008104	CLOSING:PUN-JOB HAS HELPED HIM? NO
R	523	H442000-A005/005	N008105	CLOSING:PUN-UNLOCK SOME SECRETS? YES
R	524	H442000-A006/006	N008106	CLOSING:PUN-LOT HINGES ON KINDNESS? YES
R	525	H442000-A007/007	N008107	CLOSING:MAIN PURPOSE-REPT SWEENEY LEAVES
R	526	H442000-A008/008	N008108	CLOSING:TONE OF CAPTION IS CLEVER AND WIT
R	527	H443000-A001/001	N007501	TRAVELS:MAN AFRAID-FEARFUL THOUGHTS,NO DA
R	528	H461000-A002/002		LETTER TO NANCY:WRITER KNOWS RELATIONSHIP
R	529	H461000-A003/003		LETTER TO NANCY:WRITER FEELS REJECTED AND
R	530	H463000-A001/001	N013601	SWINGING/STAR:PEOPLE LIKE PIG IF LAZY AND
R	531	H463000-A002/002	N013602	SWINGING/STAR:PEOPLE SHOULD DIFFER-TRY BE
R	532	H463000-A003/003	N013603	SWINGING/STAR:LINE 4 DOESN'T RHYME WITH O
R	533	H466000-A002/002		TEEVEE:WHAT MR&MRS SPOUSE NOT KNOW-HUSBAN
R	534	H466000-A003/003		TEEVEE:MR & MRS SPOUSE NOT TALK BECAUSE W
R	535	H466000-A004/004		TEEVEE:WRITER MAKES POEM SOUND FUNNY
R	536	H468000-A001/001	N010801	ANGRY:CHILD COMES OUT WHEN FEELS BETTER
R	537	H468000-A002/002		ANGRY:CHILD IS PERSON WHO CAN DEAL WITH O
R	538	H471000-A001/001		SONNET:POET LIES TO MAINTAIN AN ILLUSION
R	539	H471000-A002/002		SONNET:BEST THEME-LOVE FULL OF PLEASING S
R	540	H471000-A003/003		SONNET:LOVE MADE OF TRUTH MEANS NEVER LIE
SS	541	H602000-A001/001	N005701	GRAPH:MOST POWER 1980,1985,2000-PETROLEUM
SS	542	H602000-A002/002	N005702	GRAPH:IN 2000,HYDROPOWER SUPPLY LESS THAN
SS	543	H602000-A003/003	N005703	GRAPH:IN 2000 NUCLEAR POWER MORE % TOTAL
SS	544	H605000-A001/001	N007101	BUS SCHED:LAST BUS IN EVENING LEAVE CITAD
SS	545	H605000-A002/002	N007102	BUS SCHED:2ND SAT AM BUS ARRIVE DOWNTOWN
SS	546	H605000-A003/003	N007103	BUS SCHED:MISS 2:35PM FROM HANCOCK WAIT T
SS	547	H605000-A004/004	N007104	BUS SCHED:LV RUSTIC WED 9:42AM ARRV DWNTW

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
SS	548	H606000-A001/001	N012801	GRAPH:SPENT MOST ON A BOOK
SS	549	H606000-A002/002	N012802	GRAPH:RECORD COST $2.50
SS	550	H606000-A003/003	N012803	GRAPH:5 ITEMS COST MORE THAN PAINTBRUSH
SS	551	H606000-A004/004	N012804	GRAPH:SPENT SAME AMOUNT ON PAINTS,BIKE PA
SS	552	H607000-A001/001		FEB CALENDAR: FEB 18 IS THURSDAY
SS	553	H607000-A002/002		FEB CALENDAR: MONDAY OCCURS 5 TIMES IN FE
SS	554	H607000-A003/003		FEB CALENDAR: FRIDAY CLOSEST TO 10TH IS 1
SS	555	H607000-A004/004		FEB CALENDAR: MON AFTER THIRD TUES IS FEB
SS	556	H621000-A001/001	N006401	TEXTS:INDEX-BEST PLACE LOCATE BULL RUN/HS
SS	557	H621000-A002/002	N006402	TEXTS:GLOSSARY-BEST PLACE FIND DELTA DEF.
SS	558	H622000-A001/001	N006201	INDEX:FIND DARIUS INFO ON PG 23
SS	559	H622000-A002/002	N006202	INDEX:FIND CUNEIFORM PRONUNCIATION
SS	560	H622000-A003/003	N006203	INDEX:1875 FRENCH CONSTITUTION INFO ON PG
SS	561	H622000-A004/004	N006204	INDEX:ALTERNATE HDG./DUTCH EAST INDIES-IN
SS	562	H622000-A005/005	N006205	INDEX:DISARMAMENT IN EASTERN EUROPE INFO
SS	563	H624000-A001/001	N006601	TABLE CONTENTS:MOST USEFUL IN AMERICAN HI
SS	564	H624000-A002/002	N006602	TABLE CONTENTS:AMERICAN INDEPENDENCE IN U
SS	565	H624000-A003/003	N006603	TABLE CONTENTS:RECONSTRUCTION AFT CIVIL W
SS	566	H624000-A004/004	N006604	TABLE CONTENTS:MAJOR TOPIC CHAP.17-HAPPEN
SS	567	H624000-A005/005	N006605	TABLE CONTENTS:MIDDLE EAST MAP,1958-1970
SS	568	H627000-A001/001	N011901	INDEX:FIND OUT ABOUT SALMON-PGS 84&85
SS	569	H627000-A002/002	N011902	INDEX:ALTERNATE INFO-RAILRDS;TRAVEL & TRA
SS	570	H627000-A003/003	N011903	INDEX:FIND MAP OF SNAKE RIVER-PG 84
SS	571	H627000-A004/004	N011904	INDEX:FIND MAP S. AMERICAN RAIN FORESTS-P
SS	572	H629000-A001/001	N012401	INDEX:ALPHA LIST OF TOPICS AND PAGE NUMBE
SS	573	H642000-A001/001	N007004	CATALOG CD:OTHER HEADING TO FIND BOOK-SIE
SS	574	H642000-A002/002	N007002	CATALOG CD:PG FOR OTHER BOOKS SAME TOPIC-
SS	575	H642000-A003/003	N007003	CATALOG CD:AUTHORS OF BOOK-COOPER & SIEDE
SS	576	H642000-A004/004	N007001	CATALOG CD:WHAT INFO GIVES LOCATION-GV 88
SS	577	H643000-A001/001	N012201	DICTIONARY:PLUME IS FEATHER
SS	578	H643000-A002/002	N012202	DICTIONARY:MORE THAN 1 PLOWMAN IS PLOWMEN
SS	579	H643000-A003/003	N012203	DICTIONARY:PLUNDER-ROB
SS	580	H643000-A004/004	N012204	DICTIONARY:PLUM-IMPORTANT WORK
SS	581	H645000-A001/001		DICTIONARY:HOW SYLLABICATE HACKBERRY/HACK
SS	582	H645000-A002/002		DICTIONARY:PLURAL OF HABITUS- HABITUS
SS	583	H645000-A003/003		DICTIONARY:ADVERB FORM-HABITUAL/HABITUALL
SS	584	H645000-A004/004		DICTIONARY:HACKMATACK-A TYPE OF TREE
SS	585	H646000-A001/001	N011601	DICTIONARY:DEFINITION TOME-A LARGE BOOK
SS	586	H646000-A002/002	N011602	DICTIONARY:TOMORROW SYLLABICATED-TO MOR R
SS	587	H646000-A003/003	N011603	DICTIONARY:PLURAL IS TONSILLECTOMIES
SS	588	H646000-A004/004	N011604	DICTIONARY:TOLERANCE IS A NOUN
SS	589	H646000-A005/005	N011605	DICTIONARY:TONIC-MAKES YOU FEEL BETTER

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
SS	590	H647000-A001/001		GUIDE WDS:PAGE 558 FOR MUSK
SS	591	H647000-A002/002		GUIDE WDS:PAGE 560 FOR MYSTERY
SS	592	H647000-A003/003		GUIDE WDS:PAGE 561 FOR NAIAD
SS	593	H647000-A004/004		GUIDE WDS:PAGE 559 FOR MUZHIK
SS	594	H647000-A005/005		GUIDE WDS:PAGE 560 FOR MYSTIC
SS	595	H650000-A001/001	N012701	ENCYCLOPEDIA:INFO ON MEXICO IN VOLUME 6
SS	596	H650000-A002/002	N012702	ENCYCLOPEDIA:INFO ON INVENTIONS OF EDISON
SS	597	H650000-A003/003	N012703	ENCYCLOPEDIA:INFO ON IOWA FARM PRODUCTS I
SS	598	H650000-A004/004	N012704	ENCYCLOPEDIA:INFO ON N.Y. RIVERS & LAKES
SS	599	H651000-A001/001	N011501	DICTIONARY:TO FIND WORD MEANING-DICTIONAR
SS	600	H652000-A001/001	N012501	ENCYCLOPEDIA:TO FIND INFO ON WHALE FOOD-E
R	601	H662000-A001/001		GALAPAGOS:COLONIZATION UNDER HUMANS ON TH
R	602	H662000-A002/002		GALAPAGOS:OCEAN AREA APRX 36,000 SQ MILES
R	603	H662000-A003/003		GALAPAGOS:LOWLANDS-CINDER WITH SHARP EDGE
R	604	H662000-A004/004		GALAPAGOS:WOODPECKER FINCH USES TOOL TO F
R	605	H662000-A005/005		GALAPAGOS:DAGGERS LEFT BY BUCCANEERS
R	606	H662000-A006/006		GALAPAGOS:JERVIS ISLAND ABOUT 5 MI. S. OF
R	607	H662000-A007/007		GALAPAGOS:RESEARCH STATION ON INDEFATIGAB
R	608	H662000-A008/008		GALAPAGOS:ISLANDS ERUPTED FROM SEA FLOOR
R	609	H662000-A009/009		GALAPAGOS:MELVILLE SAYS ROCK RODONDO LIKE
R	610	H662000-A010/010		GALAPAGOS:VILLIERS WROTE ABOUT RETRACING
R	611	H662000-A011/011		GALAPAGOS:INDEFATIGABLE NAMED AFTER SHIP
R	612	H662000-A012/012		GALAPAGOS:DARWIN THEORY-NATURAL SELECTION
R	613	H662000-A013/013		GALAPAGOS:CORMORANT CAN'T FLY
R	614	H662000-A014/014		GALAPAGOS:CALLED CROSSROAD-2 CURRENTS MEE
R	615	H662000-A015/015		GALAPAGOS:SAILORS WITH MELVILLE WERE WHAL
R	616	H662000-A016/016		GALAPAGOS:SADDLE-SHAPE SHELLS-TALL CACTI
R	617	H662000-A017/017		GALAPAGOS:COLONIZATION FAILED-STRIFE & RE
R	618	H662000-A018/018		GALAPAGOS:NARBOROUGH-MOST AWESOME
R	619	H662000-A019/019		GALAPAGOS:ENG. NAME SANTA MARIA IS CHARLE
R	620	H662000-A020/020		GALAPAGOS:4 VEGETATION ZONES
R	621	H662000-A021/021		GALAPAGOS:SUGGEST CAREFULLY MANAGE TOURIS
R	622	H662000-A022/022		GALAPAGOS:SEYCHELLES ONLY OTHER PLACE LAN
R	623	H662000-A023/023		GALAPAGOS:SOME SPECIES SURVIVED BECAUSE D
R	624	H662000-A024/024		GALAPAGOS:PASS BARRINGTON IF SAIL ACADEMY
R	625	H662000-A025/025		GALAPAGOS:IDEAL FOR VARIETY-WARM & COOL C
R	626	H662000-A026/026		GALAPAGOS:HUMAN THREAT-NEW PLANTS & ANIMA
R	627	H662000-A027/027		GALAPAGOS:WELLINGTON BEST RECENT SOURCE
R	628	H662000-A028/028		GALAPAGOS:BISHOP DISCOVERED
R	629	H662000-A029/029		GALAPAGOS:MARINE IGUANA AT ESPINOSA POINT
R	630	H662000-A030/030		GALAPAGOS:SHARP-BEAKED GROUND FINCH DRINK
SS	631	H663000-A001/001	N006001	PHONE DIR:STORES SELL MILK LISTED UNDER D

Table B-1
(continued)

List of Items Initially Considered for Trend Analysis
(R=Reading, SS=Study Skills)

Type	No.	ECS ID	ETS ID	DESCRIPTION
SS	632	H663000-A002/002	N006002	PHONE DIR:HENDRICKS MINING ON 63RD ST, 44
SS	633	H663000-A003/003	N006003	PHONE DIR:STAR TRACKER OPEN TO REPAIR MIC

Table B-2

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
4	1.45665	0.15096	0.96805	0.25239	0.21723	0.00835
7	1.62013	0.18008	-0.66107	0.08822	0.21518	0.03084
8	2.87125	0.31297	-1.05092	0.08273	0.21736	0.02874
9	1.95374	0.19194	-0.99160	0.06814	0.16949	0.03079
10	0.66239	0.12542	0.92530	0.35950	0.22381	0.02706
11	0.32097	0.03743	-0.27738	0.08125	0.17573	0.03782
15	1.72384	0.17108	-1.71880	0.13016	0.18048	0.04286
18	1.82295	0.16808	-0.94913	0.06309	0.16900	0.02823
60	1.65532	0.09796	-1.93793	0.09259	0.14230	0.03574
66	0.94869	0.11392	-2.77999	0.25983	0.18392	0.05512
67	0.91878	0.07605	-1.14013	0.05928	0.16063	0.04083
76	1.17304	0.05089	-1.29151	0.03818	0.19673	0.02802
77	1.25806	0.07384	-1.28936	0.05040	0.21647	0.03413
78	0.94913	0.09982	-1.37673	0.08868	0.17833	0.04777
79	1.45911	0.14260	-1.46350	0.09451	0.17102	0.03989
80	1.19967	0.06978	-1.99588	0.08408	0.18779	0.04679
81	1.12587	0.11218	-1.94154	0.13624	0.17235	0.04860
82	1.32448	0.14185	-1.70034	0.12522	0.22387	0.04995
83	1.39426	0.17551	-2.74479	0.30524	0.18040	0.05359
84	1.57546	0.18945	-2.55817	0.28621	0.18460	0.05342
85	0.93831	0.14278	-3.49095	0.43919	0.18192	0.05486
86	1.59383	0.08969	-1.46392	0.05524	0.13331	0.02988
87	0.86089	0.04992	-1.52366	0.05598	0.11620	0.03398
88	1.66925	0.10922	-2.21662	0.12506	0.14307	0.04181
89	1.09681	0.10441	-2.21673	0.15663	0.15438	0.05055
90	1.50519	0.13140	-1.87898	0.12709	0.14010	0.03940
91	1.14538	0.07313	-2.16925	0.10265	0.17387	0.04762
92	1.44482	0.08494	-1.97198	0.08851	0.13813	0.03672
93	1.02755	0.06403	-2.18892	0.09931	0.13485	0.04355
94	0.92444	0.04975	-1.06644	0.04211	0.17442	0.03198
97	1.21940	0.09517	-2.87148	0.18765	0.16796	0.05013
98	0.66907	0.07745	-1.38469	0.09564	0.18951	0.05364
99	1.56126	0.15996	-2.06415	0.17448	0.18542	0.05126
101	1.69359	0.14875	-1.85667	0.12591	0.37038	0.04774
102	2.64499	0.25305	-1.94163	0.20195	0.41954	0.04549
103	1.32585	0.12925	-2.40079	0.18891	0.41107	0.06195
104	1.20762	0.11853	-0.96681	0.07508	0.39818	0.03685
106	0.66741	0.14401	0.68002	0.35685	0.19696	0.03617
107	1.22871	0.08880	-0.23181	0.07365	0.12205	0.01780
108	1.24862	0.17635	0.22229	0.20459	0.12415	0.02053
112	1.43833	0.08189	-1.82431	0.07591	0.13674	0.03630
113	1.06728	0.14794	0.27592	0.19320	0.11016	0.02078

Table B-2
(continued)

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
115	0.67148	0.12101	0.08012	0.19898	0.17877	0.04343
120	0.85304	0.09532	-1.11652	0.07640	0.18592	0.04647
121	1.41027	0.14619	-1.82682	0.13735	0.20668	0.05139
122	1.00560	0.10451	-2.20773	0.16776	0.18723	0.05418
123	1.05064	0.10112	-1.64216	0.10258	0.16097	0.04505
124	0.69985	0.15599	0.57315	0.34902	0.21732	0.03902
127	1.03001	0.09766	-1.84650	0.11914	0.14967	0.04383
128	0.95646	0.09471	-1.58179	0.09985	0.16187	0.04531
130	1.17113	0.07637	-1.15664	0.04901	0.17500	0.03294
149	1.63001	0.17429	-0.33295	0.11058	0.12634	0.02115
150	1.65918	0.09793	-1.20137	0.04581	0.17605	0.02648
157	1.34995	0.11818	-1.69445	0.10373	0.15630	0.04127
159	0.99231	0.14202	0.09872	0.17424	0.14339	0.02672
160	1.96251	0.19666	-1.82268	0.15528	0.18112	0.04375
161	0.82708	0.13658	0.44922	0.24014	0.09861	0.02574
166	0.97904	0.14260	-0.08294	0.15536	0.17794	0.03267
167	1.78542	0.22306	-0.05461	0.18073	0.17162	0.01986
168	2.50099	0.24802	0.39756	0.25583	0.16633	0.01313
169	1.13560	0.18892	1.10525	0.38822	0.24602	0.01139
170	0.65379	0.06804	-0.14449	0.09662	0.13766	0.03282
171	0.82014	0.19706	1.66266	0.64085	0.17149	0.01407
174	0.95979	0.17436	1.28951	0.43469	0.20960	0.01316
175	1.25596	0.21244	1.42966	0.46366	0.13542	0.00863
178	1.11200	0.13976	0.12194	0.15558	0.08180	0.01965
189	1.33751	0.11569	-1.07488	0.05902	0.10176	0.02931
190	0.83547	0.08371	-1.97905	0.13389	0.15101	0.04585
193	1.05468	0.11985	-0.72610	0.07396	0.14852	0.03643
196	1.80469	0.22825	0.71565	0.30623	0.14841	0.01250
198	1.13646	0.10316	-1.70297	0.10388	0.13106	0.04225
199	1.25632	0.14721	-0.50009	0.09667	0.18098	0.03112
200	0.69185	0.25980	3.38179	1.68492	0.07379	0.00878
205	1.25845	0.16086	0.83819	0.26563	0.08779	0.00940
207	1.03764	0.10284	-1.47947	0.09058	0.14524	0.04375
208	1.22173	0.09646	-0.13467	0.08884	0.16783	0.01807
214	1.00078	0.10760	-1.45260	0.09642	0.19838	0.05033
216	1.12231	0.16807	0.37240	0.22812	0.12059	0.02030
219	1.30714	0.17360	-0.01111	0.16674	0.15770	0.02433
220	0.90441	0.09095	-2.41293	0.17891	0.13717	0.04662
221	1.04075	0.05843	-1.86921	0.07365	0.16039	0.04590
222	0.86623	0.09203	-1.50656	0.09917	0.18043	0.04937
223	1.73741	0.17329	-2.17127	0.19391	0.14294	0.04481

Table B-2
(continued)

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
224	1.03601	0.05628	-1.54125	0.05345	0.10274	0.03083
225	1.13881	0.06317	-1.56618	0.05642	0.11191	0.03177
243	1.11405	0.04689	-1.55683	0.04108	0.00000	0.00000
244	0.86056	0.04431	-2.13640	0.07504	0.00000	0.00000
245	1.39376	0.07364	-2.31927	0.10130	0.00000	0.00000
246	0.42236	0.06082	-5.09262	0.62155	0.15154	0.05178
247	2.47745	0.19689	-1.79059	0.16075	0.00000	0.00000
248	1.52161	0.14305	-2.45148	0.21028	0.00000	0.00000
249	1.18964	0.09125	-1.87708	0.10397	0.00000	0.00000
250	2.21517	0.16677	-1.71067	0.12910	0.00000	0.00000
252	1.14745	0.12281	-2.90807	0.25957	0.00000	0.00000
253	0.64017	0.07784	-2.76341	0.24862	0.14679	0.05031
268	0.46158	0.11309	-0.07543	0.24320	0.47724	0.05694
279	1.70130	0.11453	-2.18679	0.12713	0.16811	0.04729
336	1.07139	0.06248	-1.81221	0.07253	0.13690	0.03869
337	0.83792	0.07518	-0.85033	0.04917	0.08419	0.02737
338	1.24921	0.13382	-1.35032	0.09046	0.20965	0.04700
339	1.16341	0.11091	-0.78563	0.06026	0.10637	0.02943
340	1.25119	0.11450	-1.41784	0.08324	0.14447	0.03794
341	1.10890	0.06620	-1.99837	0.08525	0.14442	0.04108
342	1.65120	0.10490	-1.60878	0.07133	0.21153	0.03731
344	1.21641	0.11800	-1.65139	0.10876	0.18572	0.04636
347	1.09970	0.46414	2.22396	1.47631	0.22170	0.01363
348	1.33043	0.12852	-2.13280	0.16105	0.16465	0.04706
349	0.94136	0.25125	2.10038	0.85089	0.14704	0.00905
350	0.43383	0.04097	-0.05264	0.08992	0.10881	0.03428
351	0.45252	0.18486	5.70807	2.66925	0.12353	0.01099
352	2.36995	0.20085	-1.93460	0.16690	0.17220	0.04030
353	1.81794	0.19623	-0.96893	0.07681	0.24766	0.03310
355	1.46610	0.09872	-2.09186	0.10951	0.21065	0.05111
356	1.34766	0.06790	-1.92250	0.07152	0.13172	0.03504
357	1.14389	0.07424	-1.45168	0.06249	0.20633	0.04157
362	0.85569	0.11883	-0.78045	0.09223	0.28464	0.04827
363	2.09948	0.23640	0.54432	0.27365	0.12095	0.01233
365	1.65587	0.15931	-1.74025	0.12735	0.18534	0.04517
366	1.47509	0.13858	-1.95600	0.14327	0.12681	0.04001
370	0.98695	0.09441	-0.95998	0.05837	0.10667	0.03148
371	1.45610	0.08253	-0.37323	0.05326	0.11360	0.01350
372	1.45345	0.08452	-1.75548	0.07298	0.21959	0.04232
373	2.03013	0.12710	-1.93365	0.10840	0.13714	0.03284
374	1.55085	0.12617	-2.17935	0.14486	0.17586	0.04945

Table B-2
(continued)

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
375	1.34915	0.13452	0.46445	0.18176	0.20895	0.01253
377	1.04551	0.20001	0.36449	0.28720	0.18558	0.02745
378	1.69410	0.15363	-1.51611	0.09780	0.12665	0.03532
379	1.10467	0.06725	-1.37773	0.05322	0.14688	0.03455
385	1.82456	0.22375	0.53405	0.26473	0.08996	0.01158
387	1.70937	0.15263	-1.29583	0.07823	0.15965	0.03575
388	1.42104	0.12617	-0.84776	0.06042	0.12597	0.02964
389	1.25237	0.08340	-1.20106	0.05129	0.12811	0.03096
398	2.16733	0.21882	-1.61968	0.14088	0.14230	0.03626
399	1.20873	0.08408	-1.39937	0.06249	0.13607	0.03545
400	1.66343	0.12694	-1.54185	0.08238	0.18498	0.03922
401	1.95995	0.13244	-1.34619	0.06509	0.12551	0.02678
402	1.73397	0.14022	-0.93326	0.05926	0.24043	0.03013
403	1.03200	0.10047	-0.63983	0.07386	0.21860	0.03546
404	0.81103	0.06516	-1.26452	0.06362	0.15731	0.04268
405	1.10564	0.09691	-0.78831	0.06296	0.19519	0.03503
406	1.93534	0.20619	-0.62723	0.09013	0.17492	0.02754
407	1.19942	0.18510	0.08668	0.19940	0.18363	0.02676
408	0.94338	0.12559	0.15715	0.17163	0.20520	0.02732
417	1.33547	0.10657	-2.13783	0.13200	0.17656	0.05044
418	0.87379	0.14376	-0.08002	0.16922	0.20247	0.03893
419	1.60306	0.14815	-1.40450	0.08790	0.14686	0.03775
433	0.81859	0.08661	-0.02157	0.11156	0.12768	0.02655
434	1.18256	0.11698	-0.24546	0.10031	0.19407	0.02473
435	1.34429	0.14873	-0.37569	0.10544	0.17814	0.02922
438	1.92126	0.18587	-0.70323	0.07761	0.14777	0.02830
439	1.23265	0.15929	-0.38963	0.11903	0.20591	0.03473
440	1.64252	0.16220	-1.44315	0.09785	0.18021	0.04548
441	0.73215	0.07352	-1.48787	0.08994	0.15673	0.04578
442	1.15543	0.27870	1.04316	0.54183	0.14968	0.01816
443	1.18965	0.12674	-0.42123	0.09000	0.12832	0.02891
448	1.08361	0.12266	-0.86218	0.07803	0.19149	0.04513
449	0.75963	0.08890	-1.18412	0.08352	0.18537	0.05284
450	1.12612	0.12760	-0.72118	0.08038	0.17794	0.04052
451	1.51908	0.15940	-1.19059	0.08008	0.19673	0.04207
452	1.25185	0.10821	-0.72960	0.06505	0.21056	0.03182
469	1.04435	0.07976	-1.85361	0.09809	0.18646	0.05149
470	1.17287	0.09262	-0.90836	0.05645	0.17833	0.03448
471	0.97865	0.08060	-0.72898	0.05748	0.15825	0.03284
472	1.30829	0.13259	-0.78289	0.07138	0.17652	0.03488
473	2.43421	0.24895	-1.38075	0.11377	0.20367	0.03706

Table B-2
(continued)

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
474	1.34128	0.14595	-0.49917	0.09335	0.17519	0.03115
475	1.43855	0.10385	-1.34763	0.06436	0.18225	0.03798
476	0.89898	0.14535	0.00572	0.18115	0.20719	0.03544
500	1.41644	0.14586	-1.13239	0.07708	0.18685	0.04471
501	1.31145	0.11815	-1.37300	0.07896	0.13340	0.03696
502	1.04475	0.10269	-1.03543	0.06383	0.13807	0.03753
503	1.53340	0.14219	-1.15922	0.06875	0.14062	0.03465
504	1.02968	0.10303	-0.98596	0.06324	0.14524	0.03729
505	0.93224	0.11022	-1.22248	0.08908	0.21514	0.05387
506	1.55987	0.14574	-1.11662	0.06733	0.14629	0.03411
507	1.42387	0.24339	0.33306	0.28322	0.17263	0.02080
508	1.63450	0.17901	-0.67196	0.08493	0.19917	0.03090
509	1.11028	0.13251	-0.64978	0.08506	0.18948	0.03758
510	1.05518	0.15680	-0.09065	0.16117	0.18864	0.03251
511	1.75982	0.26286	-0.04544	0.21847	0.25424	0.02367
512	1.33454	0.19020	-0.26224	0.15239	0.27183	0.03252
513	1.02310	0.13242	-0.53178	0.09803	0.20477	0.03916
514	1.29440	0.12827	-1.12835	0.07108	0.15587	0.03855
515	1.23656	0.13764	-0.28220	0.10436	0.10905	0.02429
530	1.16728	0.12330	-1.29504	0.08520	0.19283	0.04778
531	0.51957	0.11242	0.70361	0.35092	0.16853	0.04218
532	1.26473	0.13073	-1.20018	0.07778	0.18034	0.04382
536	0.85027	0.08372	-0.74637	0.06895	0.19505	0.04092
537	1.13481	0.14134	-0.36559	0.10900	0.15987	0.03232
636	1.12831	0.12857	-2.23942	0.18994	0.18080	0.05404
640	1.27522	0.15208	-0.53147	0.09437	0.16481	0.03337
643	1.10518	0.11147	-1.57629	0.10394	0.17985	0.05068
647	0.69760	0.11023	-0.16664	0.14829	0.19070	0.04640
648	0.59160	0.07916	-0.62442	0.07977	0.15105	0.04815
649	0.68113	0.23487	2.40683	1.16716	0.20290	0.02061
650	0.96024	0.14577	0.08063	0.18609	0.17753	0.03365
651	0.52144	0.10399	0.02906	0.20795	0.22364	0.05564
652	0.86335	0.30370	2.17909	1.14081	0.16658	0.01693
654	1.64918	0.23320	-0.05076	0.19829	0.21072	0.02435
655	1.04616	0.10684	-1.54971	0.10161	0.17895	0.05094
656	0.91008	0.12226	-0.52723	0.10335	0.20468	0.04590
657	0.97771	0.10406	-1.22482	0.08104	0.17939	0.04940
658	1.11430	0.12583	-0.92637	0.07852	0.19640	0.04653
659	1.85222	0.19199	-1.76251	0.15095	0.17189	0.04698
660	0.86353	0.10333	-2.34647	0.20144	0.17788	0.05367
661	1.20671	0.14872	-0.75734	0.08486	0.20772	0.04083

Table B-2
(continued)

Item Parameter Estimates and Standard Errors
Age 9 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
662	1.80891	0.19489	-1.82219	0.16094	0.18721	0.05083
663	1.84316	0.21016	-0.49321	0.11317	0.20118	0.02834
664	1.56465	0.18873	-0.47706	0.11480	0.22790	0.03125
665	1.79408	0.19723	-1.91959	0.17800	0.18332	0.05134
666	1.37500	0.14311	-1.14685	0.07887	0.19433	0.04550

Table B-3

Item Parameter Estimates and Standard Errors
Age 13 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
1	1.59669	0.09324	-0.79312	0.06893	0.16047	0.03839
2	1.56585	0.06725	-0.72331	0.04851	0.00000	0.00000
4	1.02543	0.07970	0.45207	0.06063	0.18086	0.02679
5	1.00796	0.11719	-0.00003	0.06423	0.17561	0.04379
6	0.85429	0.05997	-0.33567	0.05252	0.18056	0.04105
10	0.97733	0.10085	0.45979	0.07949	0.27963	0.03209
11	0.53348	0.05121	-0.52690	0.08108	0.22274	0.05446
12	2.28036	0.14199	1.32833	0.13528	0.10548	0.00713
13	1.05428	0.14578	1.36803	0.21623	0.28201	0.01852
14	1.26094	0.16461	1.68690	0.25396	0.14184	0.01102
16	1.58373	0.09897	-1.02675	0.08995	0.18467	0.04394
17	1.54804	0.09796	-0.38620	0.05018	0.21044	0.03485
19	0.68727	0.05280	-0.75600	0.08050	0.20467	0.05125
20	1.12722	0.13906	-1.82493	0.26332	0.22530	0.05961
21	2.07925	0.24649	-1.33136	0.25648	0.20114	0.05075
22	1.03389	0.10512	-1.19468	0.14785	0.18536	0.04971
51	0.77897	0.25343	2.29684	0.76200	0.24235	0.02525
52	0.56916	0.12481	1.33634	0.29640	0.21088	0.04429
53	1.56437	0.19188	0.40827	0.09965	0.25364	0.03077
54	1.06172	0.12614	0.28327	0.07264	0.18126	0.03697
55	1.98289	0.19297	0.07802	0.06055	0.15424	0.02918
56	1.80407	0.21783	0.35885	0.09821	0.24294	0.03041
57	0.75765	0.18116	1.57444	0.37991	0.15702	0.03079
58	0.88345	0.35001	3.04480	1.24221	0.14331	0.01460
59	0.70439	0.18681	1.84646	0.48916	0.14683	0.03016
61	0.77015	0.09384	-1.57914	0.21587	0.22203	0.05905
65	0.95088	0.13330	0.53724	0.10693	0.19689	0.03825
66	0.91527	0.15786	-2.80040	0.52411	0.22035	0.05939
69	1.70230	0.11925	-0.66598	0.07883	0.00000	0.00000
70	0.96963	0.07606	-0.06109	0.03339	0.00000	0.00000
71	0.80993	0.07230	0.38087	0.03879	0.00000	0.00000
72	1.06220	0.08377	0.34509	0.03654	0.00000	0.00000
73	1.54911	0.10640	-0.38481	0.05335	0.00000	0.00000
74	1.61040	0.10880	0.00975	0.03461	0.00000	0.00000
75	2.18640	0.14396	0.15672	0.03674	0.00000	0.00000
76	1.17766	0.05648	-1.11702	0.06996	0.22649	0.04927
92	0.94269	0.07555	-1.89281	0.17054	0.20516	0.05472
94	0.99176	0.05087	-1.16519	0.07705	0.24705	0.05469
96	1.14800	0.11845	-1.54394	0.19134	0.13873	0.04795
98	1.20288	0.14508	-1.09817	0.16472	0.27872	0.06557
99	0.84890	0.14715	-2.66391	0.49495	0.21940	0.05876
106	1.20270	0.12138	-0.15590	0.06050	0.14699	0.03857
111	0.92349	0.11275	0.59882	0.09206	0.12342	0.03099
113	1.19757	0.14462	0.33611	0.08115	0.19502	0.03591

Table B-3
(continued)

Item Parameter Estimates and Standard Errors
Age 13 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
116	1.53451	0.09436	-1.11160	0.09251	0.13645	0.04295
118	1.45582	0.08787	-0.46581	0.05068	0.16973	0.03572
119	0.58817	0.05523	-2.20859	0.22257	0.21818	0.05866
121	1.34064	0.16727	-1.79824	0.27852	0.21453	0.05699
124	0.88448	0.10726	-0.17137	0.07267	0.18722	0.05080
126	1.19999	0.13285	-0.01589	0.06935	0.21113	0.04377
131	1.28959	0.12507	-1.10239	0.13984	0.18453	0.04791
132	0.68989	0.05077	0.01906	0.04015	0.10670	0.03408
133	1.53804	0.13257	0.03750	0.05660	0.24944	0.03272
134	0.37381	0.04464	-2.55330	0.32110	0.23212	0.06189
135	1.27977	0.10897	0.20932	0.06195	0.34308	0.02921
136	1.25902	0.09101	-0.30925	0.05517	0.31344	0.03631
137	1.61458	0.11286	0.46445	0.06078	0.18000	0.01786
138	1.56286	0.16256	-1.04807	0.14887	0.19909	0.05025
139	1.20676	0.10904	-0.49670	0.07370	0.12228	0.03919
140	0.98930	0.11030	-0.05248	0.06666	0.19580	0.04515
141	1.19419	0.07091	-0.53257	0.05335	0.21024	0.03931
143	1.39277	0.09653	0.09931	0.04490	0.21217	0.02646
144	0.84293	0.17586	1.07581	0.25875	0.55192	0.02809
146	0.94167	0.21365	1.98451	0.47576	0.26071	0.01745
151	1.57669	0.14607	0.23289	0.05702	0.09484	0.02371
152	1.62846	0.15165	-0.42908	0.07411	0.14280	0.03691
154	1.11360	0.10755	-0.65840	0.09082	0.13158	0.04261
155	1.33410	0.13986	-1.44459	0.19189	0.13718	0.04716
156	0.43739	0.16024	2.70862	0.98583	0.36504	0.03397
161	0.70020	0.08487	0.37928	0.06596	0.10757	0.03520
166	1.05529	0.10363	-0.67225	0.09300	0.14074	0.04572
167	1.35363	0.12715	-0.57525	0.08489	0.14139	0.04310
168	1.51210	0.14793	-0.09857	0.05987	0.15647	0.03723
169	0.73604	0.08029	0.72449	0.09719	0.17979	0.03468
170	0.49417	0.03907	-0.87892	0.08515	0.11392	0.03992
171	0.86872	0.08187	0.71826	0.08408	0.14879	0.02669
172	1.60819	0.16330	-1.29110	0.18178	0.13407	0.04516
173	0.78839	0.08013	-1.00407	0.12300	0.11848	0.04142
174	0.87028	0.08819	0.90340	0.10679	0.18969	0.02483
176	1.08070	0.39388	2.74079	1.05989	0.09458	0.01220
177	0.64669	0.21068	2.80808	0.91453	0.14116	0.02316
178	1.39032	0.13301	-0.55855	0.08444	0.14897	0.04074
180	1.67296	0.14542	-0.08458	0.05108	0.09911	0.02726
181	0.41095	0.03595	1.44278	0.12409	0.11177	0.02241
182	1.67108	0.08483	-0.68204	0.05719	0.00000	0.00000
183	1.90013	0.09132	-0.41888	0.04326	0.00000	0.00000
184	2.76375	0.12646	-0.02558	0.03067	0.00000	0.00000

Table B-3
(continued)

Item Parameter Estimates and Standard Errors
Age 13 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
185	2.51673	0.11362	0.03787	0.02811	0.00000	0.00000
186	1.89709	0.16323	0.03287	0.05674	0.31036	0.02830
187	1.52011	0.13599	0.46874	0.07497	0.19272	0.02275
188	1.64724	0.13332	0.18187	0.05443	0.21692	0.02573
193	1.95877	0.19860	-0.67533	0.11440	0.19247	0.04197
194	0.98902	0.19316	1.45942	0.30495	0.11026	0.02332
195	1.75025	0.22692	0.62868	0.13533	0.21511	0.02552
196	1.17830	0.10074	0.31874	0.05971	0.17701	0.02896
197	0.95770	0.08378	1.28597	0.12385	0.16081	0.01401
198	0.79796	0.08325	-1.25946	0.15231	0.13215	0.04615
200	1.48339	0.20282	0.88927	0.15228	0.10267	0.01875
201	1.52326	0.10658	-1.47179	0.13543	0.20132	0.05141
203	1.11234	0.13766	0.48069	0.09308	0.15485	0.03494
204	0.69822	0.08559	-0.63959	0.10917	0.21316	0.05558
210	0.98206	0.14145	0.79140	0.13379	0.11979	0.03070
212	1.37474	0.13678	0.18716	0.05211	0.10984	0.02592
213	1.02890	0.21001	1.21627	0.27989	0.18100	0.02932
216	0.91287	0.09734	-0.04536	0.05615	0.12865	0.03981
217	1.46516	0.20963	1.01553	0.19436	0.14433	0.02279
218	1.37473	0.19135	0.59180	0.12494	0.22094	0.03070
219	1.18845	0.11622	-0.34421	0.06736	0.13454	0.04031
236	1.13623	0.06069	-0.70305	0.05228	0.00000	0.00000
237	1.51939	0.07136	-0.31981	0.03408	0.00000	0.00000
238	1.44243	0.06664	0.04160	0.02351	0.00000	0.00000
239	1.30005	0.06279	-0.29244	0.03160	0.00000	0.00000
268	0.97429	0.17236	-0.29744	0.12307	0.52631	0.05551
281	0.91828	0.07311	0.30239	0.05184	0.15225	0.03035
282	0.75264	0.05448	-0.48151	0.06002	0.17341	0.04352
288	0.53747	0.11100	1.35141	0.28754	0.35969	0.03905
289	1.05311	0.13811	0.95220	0.15262	0.30673	0.02370
291	1.00488	0.09706	-0.85783	0.11279	0.38101	0.05690
292	0.74250	0.09378	-0.59024	0.12044	0.45876	0.06141
293	1.05359	0.06176	-0.95351	0.07054	0.00000	0.00000
294	0.90762	0.10346	-1.11274	0.15293	0.21196	0.05619
310	2.26877	0.12737	-0.13535	0.03868	0.26263	0.02243
311	2.28012	0.15090	-0.19776	0.04774	0.38496	0.02543
312	1.21856	0.07243	-0.89960	0.07272	0.20930	0.04733
314	1.13782	0.12077	0.36537	0.08086	0.23563	0.03512
315	0.87611	0.08706	0.53376	0.07536	0.14318	0.03156
316	0.92071	0.12754	0.13563	0.07046	0.22264	0.04453
317	1.47172	0.13992	-0.38777	0.06973	0.15558	0.03578
318	1.48431	0.11809	-1.10771	0.12291	0.00000	0.00000
319	1.75072	0.14255	-1.12547	0.14035	0.00000	0.00000

Table B-3
(continued)

Item Parameter Estimates and Standard Errors
Age 13 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
320	1.77939	0.12727	-0.72461	0.08745	0.00000	0.00000
321	1.60588	0.10477	-0.10657	0.03790	0.00000	0.00000
322	1.22556	0.09085	-0.77735	0.08055	0.00000	0.00000
323	1.46627	0.10435	-0.72216	0.07874	0.00000	0.00000
324	1.17863	0.08565	-0.27795	0.04418	0.00000	0.00000
325	1.47065	0.10204	0.11847	0.03278	0.00000	0.00000
342	1.57489	0.10997	-1.29250	0.12075	0.20590	0.04987
345	2.16828	0.20660	1.76516	0.25481	0.11550	0.00644
346	0.52442	0.14425	2.36894	0.64336	0.13807	0.03381
347	1.22683	0.17010	0.63777	0.12459	0.21836	0.03144
348	1.36178	0.18338	-1.75266	0.29022	0.21668	0.05770
349	0.94975	0.12726	1.30400	0.19120	0.15938	0.02104
350	0.49219	0.05045	0.08122	0.05370	0.14889	0.04731
351	1.42787	0.26556	2.22290	0.48163	0.10784	0.00734
353	1.42809	0.14141	-0.99913	0.13381	0.19336	0.04723
357	1.25807	0.12921	-1.20208	0.15663	0.19510	0.05099
358	1.14876	0.07895	-0.26995	0.05181	0.26647	0.03850
359	1.28674	0.14562	-1.55190	0.21743	0.22450	0.05816
362	1.14652	0.17208	-0.62576	0.14206	0.45184	0.06434
364	0.66489	0.11928	1.05190	0.19680	0.13931	0.04052
367	1.37317	0.16554	0.52966	0.10041	0.17940	0.02892
371	1.41620	0.07708	-0.64684	0.05370	0.11312	0.03153
375	1.18090	0.07538	0.18255	0.04193	0.18640	0.02509
377	0.75895	0.06892	0.27735	0.05770	0.17556	0.03810
380	1.78845	0.20040	-0.41334	0.09078	0.31159	0.04444
385	1.17292	0.08648	0.19274	0.04600	0.09796	0.02650
386	1.55764	0.09352	-0.62054	0.05889	0.18382	0.03561
404	0.53089	0.05754	-1.12245	0.14254	0.21612	0.05709
405	1.65257	0.12072	-0.63217	0.07401	0.18340	0.04005
406	1.81441	0.17436	-0.51630	0.08919	0.18921	0.04286
407	1.47959	0.13391	-0.09604	0.05581	0.15172	0.03400
408	1.36800	0.10171	-0.13247	0.04730	0.16115	0.03350
417	1.59288	0.19539	-1.96355	0.31300	0.21223	0.05709
418	1.16827	0.12611	-0.37348	0.08079	0.22413	0.05091
419	1.27557	0.15006	-1.49448	0.21595	0.20496	0.05513
433	1.42205	0.15314	-0.22194	0.07104	0.19747	0.04458
434	1.23817	0.13747	-0.24703	0.07427	0.20430	0.04824
444	1.75023	0.19138	-0.51518	0.09747	0.21248	0.04708
447	1.32592	0.19449	0.51705	0.12671	0.26183	0.03756
448	1.56265	0.17695	-0.82890	0.13435	0.22408	0.05419
449	1.00050	0.11600	-0.92096	0.13551	0.20928	0.05465
450	1.22838	0.13124	-0.64247	0.10120	0.19097	0.04887
463	1.17000	0.10417	-0.51312	0.07779	0.26692	0.05162

Table B-3
(continued)

Item Parameter Estimates and Standard Errors
Age 13 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
464	1.68338	0.15700	0.37014	0.07753	0.25551	0.02538
465	1.10275	0.09800	-0.04452	0.05580	0.20504	0.03989
466	1.14604	0.10265	0.00016	0.05624	0.20855	0.03888
467	0.82003	0.07979	0.01609	0.05747	0.19535	0.04345
468	1.28930	0.09393	-0.03405	0.04376	0.14227	0.02976
495	1.08601	0.12819	-0.10404	0.07027	0.20781	0.04788
496	1.38072	0.15302	-0.71118	0.11365	0.21254	0.05295
497	1.85707	0.18919	-0.39627	0.08170	0.18230	0.04204
634	1.81347	0.20624	-0.08198	0.07064	0.23088	0.04016
635	1.46668	0.17948	0.08281	0.07649	0.23556	0.04182
636	0.61351	0.12897	-2.98323	0.65580	0.22377	0.06017
637	1.44689	0.34546	1.81572	0.53670	0.20173	0.01776
638	0.65209	0.21529	2.61415	0.87036	0.23101	0.02888
639	0.90061	0.10987	-0.43724	0.09253	0.22134	0.05567
640	1.23404	0.13452	-0.48771	0.08955	0.20829	0.05065
641	1.12107	0.11965	-0.51220	0.08751	0.18216	0.04726
642	1.44388	0.23487	0.65943	0.16725	0.28037	0.03470
643	1.82313	0.20953	-0.94115	0.16121	0.20699	0.05223
644	1.34890	0.20456	-1.77797	0.32609	0.21545	0.05766
645	2.03424	0.23389	-0.93154	0.17113	0.19353	0.04953
646	2.59135	0.29699	-0.35925	0.09686	0.24035	0.04191
647	1.10231	0.16135	0.06980	0.09109	0.30094	0.05370
648	0.31726	0.05983	-1.32120	0.27390	0.14976	0.05223
649	0.67975	0.21187	2.20758	0.69635	0.24745	0.03272
650	1.32778	0.16397	0.05762	0.07431	0.23338	0.04415
651	0.60083	0.08678	-0.28677	0.08693	0.21222	0.05546
652	1.07488	0.15769	0.73201	0.13547	0.14448	0.03389
654	1.09902	0.12947	-0.21836	0.07480	0.21328	0.04952
655	1.57269	0.20306	-1.44357	0.24732	0.21109	0.05599

Table B-4

Item Parameter Estimates and Standard Errors
Age 17 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
5	0.59432	0.10084	0.24954	0.11481	0.25427	0.05811
6	0.87132	0.06936	-0.40147	0.09473	0.20221	0.04389
10	1.18014	0.08729	0.30022	0.05945	0.27804	0.03177
11	0.64668	0.07597	0.10939	0.10740	0.37757	0.05674
12	1.98049	0.12169	1.04973	0.07714	0.10755	0.01163
13	1.02203	0.07951	0.99909	0.07758	0.24024	0.02364
14	0.68562	0.06332	1.55313	0.11583	0.12626	0.02287
16	0.97125	0.06572	-1.59551	0.16424	0.24518	0.05540
17	1.28363	0.08264	-0.34538	0.07810	0.31034	0.03918
19	0.53958	0.03925	-0.76418	0.11087	0.23001	0.05189
20	1.63945	0.25431	-1.52437	0.46578	0.24948	0.05780
21	1.36369	0.19198	-1.51515	0.37365	0.25025	0.05845
22	0.74009	0.09171	-1.24172	0.24129	0.25336	0.05972
48	0.87526	0.10303	-0.87556	0.19122	0.25945	0.05904
49	1.29795	0.14094	-0.12048	0.11148	0.26735	0.05014
50	1.05280	0.11257	-0.42377	0.13279	0.23556	0.05228
52	0.81997	0.10518	0.48152	0.08869	0.23286	0.04640
53	1.62017	0.18426	0.46846	0.08719	0.24928	0.03478
54	0.82743	0.09986	0.43225	0.08230	0.21730	0.04391
57	0.84634	0.11760	1.17963	0.13186	0.15362	0.03268
58	0.99274	0.22421	2.28873	0.46741	0.15108	0.02084
59	0.55757	0.17453	2.65016	0.68789	0.18257	0.03061
62	1.53129	0.16289	1.06191	0.13498	0.49287	0.01982
63	2.35399	0.25031	0.81327	0.13001	0.47558	0.02091
64	3.22380	0.42294	0.88946	0.21025	0.59051	0.01772
65	0.88214	0.06660	0.00287	0.06824	0.23844	0.03966
67	1.10620	0.10584	-1.83938	0.26790	0.17537	0.05049
69	1.02922	0.09581	-1.22016	0.19053	0.00000	0.00000
70	0.44947	0.04969	0.54968	0.03421	0.00000	0.00000
71	0.65227	0.05710	0.20389	0.04841	0.00000	0.00000
72	0.97057	0.07134	0.41072	0.04013	0.00000	0.00000
73	1.16801	0.09325	-0.62411	0.11847	0.00000	0.00000
74	0.75477	0.06220	-0.08259	0.06687	0.00000	0.00000
75	1.30843	0.09016	0.05606	0.05984	0.00000	0.00000
94	0.81195	0.04418	-1.40886	0.11973	0.23908	0.05287
95	0.46229	0.06794	-1.13349	0.26065	0.21025	0.05941
96	0.84699	0.10985	-1.82007	0.33210	0.19377	0.05585
107	0.88547	0.05288	-0.65315	0.08583	0.14317	0.03849
108	0.67852	0.08229	-0.18683	0.11775	0.23676	0.05346
109	1.41434	0.17188	0.97126	0.12507	0.21527	0.02901
110	1.17945	0.17302	1.08039	0.15213	0.22717	0.03149
113	0.95333	0.10833	0.50002	0.07917	0.19986	0.03966
114	1.22417	0.11490	-0.34576	0.11349	0.14786	0.04077
115	0.84714	0.09042	-1.07396	0.19467	0.19175	0.05418

Table B-4
(continued)

Item Parameter Estimates and Standard Errors
Age 17 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
124	0.69228	0.12098	0.50749	0.11257	0.30093	0.05240
125	1.35283	0.14904	-0.95722	0.21429	0.16147	0.04570
126	1.14087	0.07039	-0.23730	0.06786	0.19640	0.03548
133	1.66698	0.12274	-0.17206	0.08357	0.18573	0.03129
134	0.36123	0.04104	-2.48841	0.35356	0.27880	0.06533
135	1.35691	0.11314	0.51104	0.06978	0.41078	0.02682
136	0.98272	0.08258	-0.44557	0.11038	0.42686	0.04811
137	1.57677	0.09791	0.16802	0.05176	0.21842	0.02553
138	1.14654	0.13888	-1.31181	0.26686	0.23303	0.05472
139	1.04763	0.11700	-0.78478	0.17508	0.21912	0.05571
140	0.79883	0.09200	-0.05900	0.10488	0.23301	0.05174
141	0.88706	0.05596	-0.92831	0.10928	0.25088	0.05185
142	0.45010	0.09018	1.77775	0.27284	0.23565	0.03793
143	0.98092	0.06374	0.03659	0.05851	0.19202	0.03446
144	0.74808	0.08066	0.41074	0.09176	0.49653	0.04012
146	0.61987	0.07859	1.52425	0.15434	0.24636	0.03045
147	1.57044	0.21814	0.11486	0.12772	0.49320	0.04714
148	1.45125	0.16767	-0.63203	0.18246	0.25910	0.05495
151	1.33127	0.13468	0.27173	0.07635	0.17931	0.03757
152	1.12274	0.12332	-1.04790	0.20798	0.19010	0.05307
153	1.96260	0.13277	1.71999	0.14176	0.06830	0.00657
156	0.46243	0.14180	4.09359	1.11079	0.28630	0.02185
162	0.60844	0.07624	-0.41443	0.14489	0.25092	0.05722
163	1.16963	0.12727	-0.07951	0.10308	0.21424	0.04755
164	0.88216	0.10423	-1.29489	0.24418	0.26669	0.06140
166	1.02472	0.10894	-0.54436	0.14116	0.19722	0.05151
167	1.10947	0.11192	-0.48091	0.13101	0.19344	0.04916
168	1.11519	0.10319	-0.15072	0.09229	0.14279	0.03897
169	0.74494	0.06133	0.85600	0.06342	0.13239	0.03076
170	0.22835	0.03620	-1.45780	0.32761	0.20260	0.05773
171	0.85246	0.10091	1.08234	0.11147	0.24024	0.03287
174	0.93405	0.06458	0.71370	0.05491	0.22718	0.02588
176	1.09899	0.15720	1.67135	0.21015	0.10532	0.01952
177	1.79719	0.25008	1.77935	0.29740	0.20325	0.01663
180	0.83532	0.09349	0.19459	0.08482	0.18657	0.04754
181	0.69930	0.06199	1.46190	0.10658	0.21872	0.02270
182	1.51833	0.10459	-1.17939	0.17023	0.00000	0.00000
183	1.51280	0.10451	-1.22430	0.17530	0.00000	0.00000
184	2.73854	0.15404	-0.83845	0.18271	0.00000	0.00000
185	2.46922	0.14869	-0.87849	0.18119	0.00000	0.00000
186	1.18504	0.10607	-1.20315	0.18815	0.25078	0.05289
187	1.26249	0.10020	0.03545	0.07130	0.24953	0.03616
188	1.05025	0.08945	-0.47839	0.10922	0.24689	0.04866

Table B-4
(continued)

Item Parameter Estimates and Standard Errors
Age 17 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
191	0.45199	0.06813	-1.58538	0.33402	0.26496	0.06224
193	1.04098	0.11131	-0.43962	0.13478	0.24884	0.05275
194	0.64105	0.12180	1.82739	0.28054	0.16678	0.03701
195	0.53332	0.08243	0.43129	0.09750	0.26520	0.05620
196	1.77071	0.14280	0.69601	0.06980	0.22798	0.02208
197	1.29400	0.06383	1.10911	0.05357	0.14971	0.01153
200	1.90838	0.17310	0.60725	0.07112	0.11621	0.02207
201	1.21404	0.08423	-1.31791	0.15585	0.22761	0.04843
202	0.96720	0.12330	-1.62836	0.31339	0.25683	0.06016
203	0.63631	0.08083	0.93656	0.09236	0.13972	0.03724
204	0.47799	0.06931	-0.41630	0.16504	0.26466	0.06129
205	0.71847	0.07537	1.09384	0.09238	0.12780	0.03100
206	0.85364	0.09833	-0.22248	0.11972	0.26638	0.05256
210	0.62594	0.08744	0.95847	0.10476	0.16588	0.04229
212	1.11070	0.11419	0.03431	0.08767	0.18187	0.04405
213	0.91789	0.10593	0.67046	0.07897	0.14137	0.03576
216	0.49083	0.06256	-0.21556	0.12128	0.18109	0.05149
217	1.26543	0.16442	1.18444	0.15028	0.17113	0.02743
236	0.69610	0.04781	-1.19105	0.12801	0.00000	0.00000
237	1.06995	0.06170	-0.60850	0.08244	0.00000	0.00000
238	1.10482	0.05601	0.22346	0.03486	0.00000	0.00000
239	0.93467	0.05132	-0.35216	0.06172	0.00000	0.00000
240	0.80081	0.06984	-0.70598	0.12303	0.00000	0.00000
241	0.59960	0.06427	-1.36050	0.21402	0.00000	0.00000
242	0.50169	0.05357	-0.70475	0.14197	0.00000	0.00000
254	0.90757	0.05137	-1.33950	0.11965	0.00000	0.00000
255	0.92322	0.09005	-1.49053	0.22328	0.00000	0.00000
256	1.70086	0.12609	-0.39853	0.11646	0.00000	0.00000
257	2.08256	0.17280	-0.55263	0.17302	0.00000	0.00000
258	1.76546	0.11580	0.13425	0.06243	0.00000	0.00000
259	1.85470	0.12134	0.19678	0.05989	0.00000	0.00000
281	1.17509	0.09228	0.42697	0.06140	0.29149	0.02980
282	0.57491	0.04444	-0.67859	0.10987	0.21371	0.04855
288	0.66453	0.08966	1.34667	0.15538	0.38141	0.03358
289	0.69711	0.07315	0.81770	0.08427	0.26561	0.03784
293	1.00109	0.09872	-2.33080	0.32026	0.00000	0.00000
294	1.55453	0.24024	-2.09024	0.57388	0.25189	0.05931
310	1.97597	0.10814	-0.20875	0.06681	0.22230	0.02815
311	1.83938	0.10728	-0.41236	0.08265	0.23967	0.03295
312	1.06266	0.06358	-1.07107	0.11463	0.22434	0.04848
314	0.83053	0.06001	0.11310	0.05968	0.20215	0.03653
315	0.99330	0.07254	0.63841	0.05335	0.18550	0.02636
316	0.46904	0.07085	0.21633	0.10438	0.25614	0.05750

Table B-4
(continued)

Item Parameter Estimates and Standard Errors
Age 17 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
317	0.99313	0.12426	0.05050	0.10744	0.28210	0.05110
318	1.08016	0.10370	-1.46409	0.23019	0.00000	0.00000
319	1.11911	0.12121	-1.66604	0.28882	0.00000	0.00000
320	1.22579	0.09701	-0.67678	0.12927	0.00000	0.00000
321	1.14684	0.07858	-0.12357	0.06981	0.00000	0.00000
322	1.35538	0.12050	-1.03931	0.18702	0.00000	0.00000
323	1.33168	0.12319	-1.09411	0.19858	0.00000	0.00000
324	1.06698	0.08468	-0.64108	0.11602	0.00000	0.00000
325	1.15749	0.07953	-0.00904	0.06058	0.00000	0.00000
339	0.62174	0.07576	-1.51617	0.26507	0.18479	0.05327
345	1.48303	0.10966	1.47069	0.10829	0.07820	0.01010
346	0.80011	0.12713	1.43191	0.18972	0.18799	0.03564
347	0.85817	0.11307	0.00737	0.11694	0.30546	0.05666
349	1.16557	0.08068	1.05905	0.07224	0.17184	0.01921
350	0.36541	0.03532	0.59877	0.05512	0.18704	0.03891
351	1.27585	0.08354	1.60148	0.10023	0.06598	0.00967
354	0.69296	0.10069	0.75584	0.10257	0.21238	0.04809
358	1.01825	0.06307	-0.23213	0.06843	0.25487	0.03836
360	0.95496	0.15098	-2.17454	0.47425	0.26545	0.06258
363	1.72319	0.16888	-0.02255	0.09748	0.19835	0.03719
364	0.72257	0.05923	0.77986	0.06076	0.14067	0.03097
367	0.61159	0.04925	0.30881	0.05634	0.17399	0.03693
375	1.06069	0.06546	0.13339	0.05284	0.22482	0.03233
377	0.48717	0.03950	-0.23439	0.08782	0.22231	0.04993
381	1.12524	0.12230	-0.13496	0.10626	0.21654	0.04634
385	1.22842	0.09807	0.18844	0.06306	0.21048	0.03479
386	1.11436	0.08874	-0.45880	0.10243	0.23585	0.04671
390	1.96132	0.27034	0.83773	0.16045	0.48969	0.02547
391	1.97781	0.21314	0.66570	0.09249	0.25281	0.02494
392	2.09415	0.26210	0.55245	0.12213	0.55618	0.02999
393	1.96949	0.33926	1.45088	0.33080	0.50612	0.01939
433	0.99114	0.11028	-0.33099	0.12683	0.25532	0.05669
434	1.22338	0.13914	-0.24237	0.12320	0.26516	0.05549
444	1.32800	0.16288	-0.92151	0.22347	0.25819	0.05934
447	1.08191	0.12525	-0.45185	0.14587	0.28064	0.06060
463	1.07046	0.10094	-0.50005	0.12829	0.39462	0.05974
464	1.53342	0.15036	0.84850	0.10722	0.42746	0.02443
465	2.16259	0.20103	0.46049	0.07854	0.45312	0.02449
466	1.20625	0.08530	0.06940	0.06193	0.19592	0.03371
467	0.72010	0.05933	0.28322	0.06108	0.18267	0.03798
468	1.35830	0.08374	0.09397	0.05236	0.12755	0.02514
495	0.94583	0.10372	-0.14280	0.10995	0.24283	0.05509
496	1.39056	0.15146	-0.51094	0.15331	0.23921	0.05392

Table B-4
(continued)

Item Parameter Estimates and Standard Errors
Age 17 Trend Data

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
497	1.09546	0.12542	-0.67282	0.16884	0.25740	0.05886
516	3.31719	0.29666	-0.32297	0.20806	0.14302	0.02464
517	1.46078	0.13492	-0.25521	0.10934	0.18331	0.03473
518	1.99442	0.18488	0.08683	0.08734	0.13598	0.02394
634	0.85589	0.10233	-0.65988	0.16776	0.10233	-0.65988
635	0.95778	0.10611	-0.14893	0.11167	0.25762	0.05682
636	0.44915	0.10409	-3.69624	0.99511	0.27166	0.06395
637	1.42735	0.17687	1.11024	0.15315	0.22818	0.03079
638	0.83278	0.18982	1.89726	0.37568	0.28629	0.03509
639	1.06417	0.12958	-0.19962	0.12797	0.30821	0.06005
640	1.04213	0.11522	-0.57695	0.15061	0.24451	0.05578
641	1.12321	0.13470	-0.84345	0.19809	0.26560	0.06046
642	1.01585	0.11815	0.55728	0.09316	0.23903	0.04710
643	1.45430	0.20012	-1.12326	0.29676	0.26229	0.06082
644	1.70941	0.32433	-1.58179	0.58078	0.25611	0.06041
645	1.38063	0.17732	-1.04481	0.25864	0.24647	0.05768
646	2.22505	0.26135	-0.33018	0.18219	0.26073	0.05078
647	1.23280	0.14713	0.22322	0.10538	0.33978	0.05467
648	0.23808	0.04790	-1.52518	0.42730	0.20636	0.05911
649	0.60196	0.13288	1.65869	0.29578	0.28403	0.04757
650	1.35563	0.14548	-0.02791	0.10287	0.27185	0.05205
651	1.02332	0.11280	-0.09762	0.10755	0.26029	0.05557
652	1.20426	0.12332	0.55375	0.07724	0.17273	0.03877
653	1.41871	0.15956	0.87244	0.10998	0.21576	0.03286
654	1.11849	0.12182	-0.36268	0.13094	0.26608	0.05827
655	1.31916	0.21214	-1.59599	0.42450	0.26037	0.06143

Table B-5
IRT TREND ANALYSIS ITEM TABLE FOR AGE 9

POS	ECS ID	\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	6	7	9	1	2	3	1	2	3	4	10	11	1	2	3	4	YEARS
		NAEP ID\																						USED
4	7099002-A002/002	N001802						14				13			13								15	4
7	7099005-A001/001		7																					1
8	7099005-A002/002		7																					1
9	7099005-A003/003		7																					1
10	7099006-A001/001					11						4			4									3
11	7099006-A002/002					11						4			4									3
15	7099008-A001/001			7																				1
18	7099011-A001/001						4																	1
60	7099024-A001/001									3			3			2								3
66	7101007-A001/001				10																			1
67	7101009-A001/001				17								2											2
76	7101055-A001/001	N004101	5								22	19	19	22	19	18							16	4
77	7101056-A001/001	N014001		10							8			8							14			4
78	7101057-A001/001				12																			1
79	7101058-A001/001					2																		1
80	7101059-A001/001	N009101						6				3			3							18		4
81	7101060-A001/001								8															1
82	7101061-A001/001									10														1
83	7101062-A001/001		11																					1
84	7101063-A001/001									5														1
85	7101064-A001/001								1															1
86	7102001-A001/001		13								19			19										3
87	7102004-A001/001				15						13			13										3
88	7102005-A001/001					1						12			12									3
89	7102006-A001/001						2																	1
90	7102007-A001/001						6																	1
91	7102008-A001/001							10				16			16									3
92	7102010-A001/001								5				5			4								3
93	7102011-A001/001								9				12			11								3
94	7102013-A001/001										20	17	17	20	17	16								2
97	7102029-A001/001		2									1			1									3
98	7102030-A001/001			11																				1
99	7102031-A001/001				2																			1
101	7102032-A002/002						15					5												2
102	7102032-A003/003						15					5												2
103	7102032-A004/004						15					5												2
104	7102032-A005/005						15					5												2
106	7102034-A001/001			15																				1
107	7102035-A001/001								17			15			15									3
108	7102036-A001/001					14																		1
112	7103002-A001/001			1							11			11										3
113	7103004-A001/001			9																				1
115	7103017-A001/001		10																					1
120	7103026-A001/001			6																				1
121	7103027-A001/001				16																			1
122	7103028-A001/001					6																		1
123	7103029-A001/001							11																1
124	7103030-A001/001								10															1
127	7103033-A001/001								16															1

Table B-5

IRT TREND ANALYSIS ITEM TABLE FOR AGE 9

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	6	7	9	1	2	3	1	2	3	4	10	11	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																						USED
128	7103034-A001/001									6														1
130	7103036-A001/001							2					4			3								3
149	7103047-A001/001					15																		1
150	7103048-A001/001								12				9			8								3
157	7103053-A001/001						13																	1
159	7103055-A001/001					4																		1
160	7103056-A001/001						1																	1
161	7103057-A001/001								15															1
166	7103062-A001/001			12																				1
167	7103062-A002/002			12																				1
168	7103062-A003/003			12																				1
169	7127001-A001/001	N003801								14			16			15							[illegible]	4
170	7127001-A002/002									14			16			15								3
171	7127001-A003/003									14			16			15								3
174	7127003-A001/001	N002101			11						15			15									5	4
175	7127003-A002/002				11						15			15										3
178	7127005-A001/001								6															1
189	7201002-A001/001				1																			1
190	7201003-A001/001							5																1
193	7201023-A001/001		14																					1
196	7202003-A001/001						7																	1
198	7203002-A001/001				3																			1
199	7203003-A001/001					8																		1
200	7203006-A001/001						14																	1
205	7203012-A001/001								7				6			5								3
207	7203043-A001/001		3																					1
208	7203044-A001/001					13					6			6										3
214	7203050-A001/001							9																1
216	7203052-A001/001									11														1
219	7227001-A001/001				7																			1
220	7301002-A001/001			3																				1
221	7301004-A001/001	N009601			5						3			3						8				4
222	7301006-A001/001					7																		1
223	7301007-A001/001						8																	1
224	7301011-A001/001							16				11			11									3
225	7301012-A001/001								13				14			13								3
243	7301071-A001/001						18				7			7										3
244	7301071-A002/002						18				7			7										3
245	7301071-A003/003						18				7			7										3
246	7302001-A001/001		8										1			1								3
247	7302002-A001/002		17																					1
248	7302002-A003/004		17																					1
249	7302002-A005/006		17																					1
250	7302002-A007/008		17																					1
252	7302004-A002/002			16																				1
253	7302005-A001/001				9																			1
263	7303004-A001/001					10																		1
279	7303013-A001/001									1			8			7								3
336	7401001-A001/001		1								5			5										[illegible]
337	7401003-A001/001			8																				1

Table B-5

IRT TREND ANALYSIS ITEM TABLE FOR AGE 9

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	6	7	9	1	2	3	1	2	3	4	10	11	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																						USED
338	7401005-A001/001		.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
339	7401007-A001/001		.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
340	7401010-A001/001		.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
341	7401011-A001/001		.	.	.	12	.	.	.	.	.	8	.	.	8	.	.	.	.	.	.	.	.	3
342	7401016-A001/001		.	.	.	.	.	17	.	.	.	2	.	.	2	.	.	.	.	.	.	.	.	3
344	7401022-A001/001		.	.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	1
347	7401032-A001/001		.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
348	7401066-A001/001		.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
349	7401067-A001/001	N003001	.	5	.	.	.	.	.	.	10	.	.	10	.	.	.	.	.	.	.	14	.	4
350	7401067-A002/002	N003002	.	5	.	.	.	.	.	.	10	.	.	10	.	.	.	.	.	.	.	15	.	4
351	7401067-A003/003	N003003	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	.	2
352	7401068-A001/001		.	.	.	.	.	.	.	.	.	.	11	.	.	10	.	.	.	.	.	.	.	2
353	7401069-A001/001		.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	1
355	7401071-A001/001		.	14	.	.	.	.	.	.	17	.	.	17	.	.	.	.	.	.	.	.	.	3
356	7401072-A001/001	N013301	.	.	.	3	.	.	.	.	.	14	.	.	14	.	.	.	.	14	.	.	.	4
357	7401073-A001/001		.	.	.	.	12	.	.	.	9	.	.	9	.	.	.	.	.	.	.	.	.	3
362	7401078-A001/001		.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
363	7401079-A001/001		.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	1
365	7401081-A001/001		.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
366	7401082-A001/001		.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
370	7401086-A001/001		.	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	1
371	7402020-A001/001	N002401	.	.	.	.	.	.	.	9	.	.	13	.	.	12	.	.	.	.	7	.	.	4
372	7402021-A001/001	N009201	.	.	14	.	.	.	.	.	16	.	.	16	.	.	.	.	.	.	.	28	.	4
373	7402022-A001/001		.	.	.	.	.	3	.	.	.	9	.	.	9	.	.	.	.	.	.	.	.	3
374	7402023-A001/001		.	.	.	.	.	.	.	.	2	.	.	2	.	.	.	.	.	.	.	.	.	2
375	7403007-A001/001	N004901	12	.	.	.	.	.	.	.	.	7	.	.	7	.	.	.	.	.	.	17	.	4
377	7403019-A001/001	N002501	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	1
378	7502012-A001/001		.	.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
379	7503001-A001/001		.	.	.	.	.	8	.	.	.	.	7	.	.	6	.	.	.	.	.	.	.	3
385	7503044-A001/001		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	.	.	1
387	H201000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	1
388	H201000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	1
389	H201000-A003/003	N008603	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	17	.	.	2
398	H205000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	1
399	H205000-A002/002	N010502	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	4	.	.	2
400	H205000-A003/003	N010503	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	5	.	.	2
401	H205000-A004/004	N010504	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	6	.	.	2
402	H206000-A001/001	N011301	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	21	.	2
403	H206000-A002/002	N011302	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	22	.	2
404	H222000-A001/001	N001601	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	8	.	2
405	H222000-A002/002	N001602	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	9	.	2
406	H222000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	1
407	H222000-A004/004		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	1
408	H222000-A005/005	N001604	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	11	.	2
417	H241000-A001/001	N004401	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	9	2
418	H241000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	1
419	H241000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	1
433	H265000-A001/001	N002001	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	10	.	.	2
434	H265000-A002/002	N002002	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	11	.	.	2
435	H265000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	1

Table B-5

IRT TREND ANALYSIS ITEM TABLE FOR AGE 9

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2 1	2 2	2 3	2 4	2 5	2 6	2 7	2 9	6 1	6 2	6 3	11 1	11 2	11 3	11 4	11 10	11 11	15 1	15 2	15 3	15 4	NO. YEARS USED
438	H268000-A001/001	N013201	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	1
439	H269000-A002/002	N010102	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	1
440	H269000-A003/003	N010103	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	1
441	H282000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	1
442	H282000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	1
443	H282000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	1
448	H286000-A001/001	N004701	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	.	.	.	1
449	H286000-A002/002	N004702	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	.	.	.	1
450	H286000-A003/003	N004703	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.	1
451	H287000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	1
452	H287000-A002/002	N013502	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	8	2
469	H404000-A001/001	N013101	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	1	.	.	.	2
470	H404000-A002/002	N013102	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	2	.	.	.	2
471	H404000-A003/003	N013103	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	3	.	.	.	2
472	H404000-A004/004		.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	1
473	H405000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	1
474	H405000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	1
475	H405000-A003/003	N001503	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	3	.	2
476	H405000-A004/004		.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	1
500	H416000-A003/003	N010003	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	1
501	H417000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	1
502	H417000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	1
503	H417000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	1
504	H417000-A004/004		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	1
505	H418000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	1
506	H418000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	1
507	H418000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	1
508	H419000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	.	1
509	H419000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	.	1
510	H419000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	.	1
511	H419000-A004/004		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	.	1
512	H419000-A005/005		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	.	1
513	H422000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	1
514	H422000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	1
515	H422000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	1
530	H463000-A001/001		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	1
531	H463000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	1
532	H463000-A003/003		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	1
536	H468000-A001/001	N010801	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	29	.	2
537	H468000-A002/002		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	1
636		N005101	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	1
640		N002003	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	1
643		N004801	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	19	.	1
647		N001603	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	1
648		N003802	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	1
649		N003803	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	1
650		N004201	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	1
651		N004202	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	1
652		N002102	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	1
654		N004402	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	1

Table B-5

IRT TREND ANALYSIS ITEM TABLE FOR AGE 9

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	6	7	9	1	2	3	1	2	3	4	10	11	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																						USED
655		N004403	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	1
656		N013104	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	1
657		N010002	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	1
658		N011101	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	36	.	.	.	1
659		N010501	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	1
660		N010301	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	1
661		N008602	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	.	.	1
662		N001501	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1
663		N001502	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	1
664		N001504	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	1
665		N010201	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	1
666		N013501	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	1
NUMBER OF CALIBRATED ITEMS IN BOOKLET			17	19	16	15	17	12	14	15	20	20	18	20	16	17	17	20	16	14	14	19	15	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET ACROSS YEARS			6	6	6	6	8	8	6	6	20	20	18	20	16	17	7	6	5	5	8	13	6	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET WITHIN YEARS			0	0	0	0	0	0	0	0	2	2	2	2	2	2	0	0	0	0	0	0	0	

Table B-6

IRT TREND ANALYSIS ITEM TABLE FOR AGE 13

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO.
			1	2	3	4	5	11	12	13	1	2	3	1	2	3	14	1	2	3	4	YEARS USED
1	7099001-A001/001									3	17			17								3
2	7099001-A002/003									3	17			17								3
4	7099002-A002/002	N001802			8								7			8					21	4
5	7099003-A001/001		3																			1
6	7099004-A001/001							12				12			13							3
10	7099006-A001/001								7				2			2						3
11	7099006-A002/002								7				2			2						3
12	7099007-A001/001				13								5			5						3
13	7099007-A002/002	N005002			13								5			5			7			4
14	7099007-A003/003	N005003			13								5			5			8			4
16	7099009-A001/001	N003601										13			14			13				3
17	7099009-A002/002	N003602										13			14			14				3
19	7099012-A001/001	N003501										2			2				10			3
20	7099013-A001/001		4																			1
21	7099013-A002/002		4																			1
22	7099013-A003/003		4																			1
51	7099021-A001/001		9																			1
52	7099021-A002/002		9																			1
53	7099021-A003/003		9																			1
54	7099021-A004/004		9																			1
55	7099022-A001/001							13														1
56	7099022-A002/002							13														1
57	7099023-A001/001								6													1
58	7099023-A002/002								6													1
59	7099023-A003/003								6													1
61	7099025-A001/001		1																			1
65	7099027-A001/001			11																		1
66	7101007-A001/001								9													1
69	7101019-A001/002								18													1
70	7101019-A003/004								18													1
71	7101019-A005/006								18													1
72	7101019-A007/008								18													1
73	7101019-A009/010								18													1
74	7101019-A011/012								18													1
75	7101019-A013/014								18													1
76	7101055-A001/001	N004101									20	20	17	20	20	16					22	3
92	7102010-A001/001				3								4			4						3
94	7102013-A001/001							1			18	18	15	18	18	14						3
96	7102015-A001/001								8													1
98	7102030-A001/001								1													1
99	7102031-A001/001				15																	1
106	7102034-A001/001			13																		1
111	7102038-A001/001							4														1
113	7103004-A001/001							2														1
116	7103019-A001/001									11	13			14								3
118	7103021-A001/001									6	11			12								3
119	7103025-A001/001			1							1			1								3
121	7103027-A001/001					5																1
124	7103030-A001/001			2																		1

Table B-6

IRT TREND ANALYSIS ITEM TABLE FOR AGE 13

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	11	12	13	1	2	3	1	2	3	14	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																				USED
126	7103032-A001/001				1																	1
131	7103037-A001/001			4																		1
132	7103038-A001/001						2					4			4							3
133	7103039-A001/001										7			8								2
134	7103041-A001/001			9							3			3								3
135	7103041-A002/002			9							3			3								3
136	7103041-A003/003			9							3			3								3
137	7103041-A004/004			9							3			3								3
138	7103042-A001/001				4																	1
139	7103042-A002/002				4																	1
140	7103043-A001/001				9																	1
141	7103044-A001/001	N001701				11					10			11					3			4
143	7103044-A003/003					11					10			11								3
144	7103045-A001/001	N005201										10			11						23	3
146	7103045-A003/003	N005203										10			11						25	3
151	7103049-A001/001					6																1
152	7103049-A002/002					6																1
154	7103051-A001/001						5															1
155	7103051-A002/002						5															1
156	7103052-A001/001						7															1
161	7103057-A001/001		8																			1
166	7103062-A001/001									14												1
167	7103062-A002/002									14												1
168	7103062-A003/003									14												1
169	7127001-A001/001	N003801	13										10			11					2	4
170	7127001-A002/002		13										10			11						3
171	7127001-A003/003		13										10			11						3
172	7127002-A001/001					12																1
173	7127002-A002/002					12																1
174	7127003-A001/001	N002101								13	5			6							5	4
176	7127004-A001/001									9												1
177	7127004-A002/002									9												1
178	7127005-A001/001			14																		1
180	7127006-A001/001						6															1
181	7127007-A001/001								3		19	19	16	19	19	15						3
182	7127009-A001/002								16				13									2
183	7127009-A003/004								16				13									2
184	7127009-A005/005								16				13									2
185	7127009-A006/006								16				13									2
186	7127009-A007/007								16				13									2
187	7127009-A008/008								16				13									2
188	7127009-A009/009								16				13									2
193	7201023-A001/001						9															1
194	7201024-A001/001					4																1
195	7201025-A001/001							8														1
196	7202003-A001/001											7			8							2
197	7202008-A001/001			5							21	21	18	21	21	17						3
198	7203002-A001/001		11																			1
200	7203006-A001/001								14													1
201	7203009-A001/001							9				6			7							3

Table B-6
IRT TREND ANALYSIS ITEM TABLE FOR AGE 13

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2 1	2 2	2 3	2 4	2 5	2 11	2 12	2 13	6 1	6 2	6 3	11 1	11 2	11 3	11 14	15 1	15 2	15 3	15 4	NO. YEARS USED
203	7203010-A001/001		.	.	.	.	.	.	.	12	.	.	.	.	.	.	.	.	.	.	.	1
204	7203011-A001/001		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	1
210	7203046-A001/001		.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
212	7203048-A001/001		.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	1
213	7203049-A001/001		.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	1
216	7203052-A001/001		.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	1
217	7203053-A001/001		.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
218	7203054-A001/001		.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
219	7227001-A001/001		.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
236	7301022-A001/002		.	.	.	.	.	16	.	.	.	17	.	.	.	.	.	.	.	.	.	2
237	7301022-A003/004		.	.	.	.	.	16	.	.	.	17	.	.	.	.	.	.	.	.	.	2
238	7301022-A005/006		.	.	.	.	.	16	.	.	.	17	.	.	.	.	.	.	.	.	.	2
239	7301022-A009/010		.	.	.	.	.	16	.	.	.	17	.	.	.	.	.	.	.	.	.	2
268	7303004-A001/001		.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	1
281	7303017-A001/001		6	.	.	.	.	.	.	.	.	.	11	.	.	12	.	.	.	.	.	3
282	7303017-A002/002		6	.	.	.	.	.	.	.	.	.	11	.	.	12	.	.	.	.	.	3
288	7303019-A001/001	N001201	.	6	.	.	.	.	.	.	2	.	.	2	.	.	.	.	.	26	.	4
289	7303019-A002/002		.	6	.	.	.	.	.	.	2	.	.	2	.	.	.	.	.	.	.	3
291	7303026-A001/001		.	.	.	15	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	2
292	7303026-A002/002		.	.	.	15	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	2
293	7303026-A003/004		.	.	.	15	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	2
294	7303026-A005/005		.	.	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	1
310	7303051-A001/001	N002201	.	.	7	.	.	.	.	.	.	.	6	.	.	7	.	.	14	.	.	4
311	7303051-A002/002	N002202	.	.	7	.	.	.	.	.	.	.	6	.	.	7	.	.	15	.	.	4
312	7303051-A003/003	N002203	.	.	7	.	.	.	.	.	.	.	6	.	.	7	.	.	16	.	.	4
314	7303054-A001/001		.	.	.	.	.	.	.	.	.	.	9	.	.	10	.	.	.	.	.	2
315	7303054-A002/002		.	.	.	.	.	.	.	.	.	.	9	.	.	10	.	.	.	.	.	2
316	7303055-A002/002		.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	1
317	7303055-A003/003		.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	1
318	7303056-A001/002		.	.	.	.	.	18	.	.	.	.	.	.	.	.	.	.	.	.	.	1
319	7303056-A003/004		.	.	.	.	.	18	.	.	.	.	.	.	.	.	.	.	.	.	.	1
320	7303056-A005/006		.	.	.	.	.	18	.	.	.	.	.	.	.	.	.	.	.	.	.	1
321	7303056-A007/008		.	.	.	.	.	18	.	.	.	.	.	.	.	.	.	.	.	.	.	1
322	7303057-A001/002		.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	1
323	7303057-A003/004		.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	1
324	7303057-A005/006		.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	1
325	7303057-A007/008		.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	1
342	7401016-A001/001		.	10	.	.	.	.	.	.	8	.	.	9	.	.	.	.	.	.	.	3
345	7401024-A001/001		.	7	.	.	.	.	.	.	4	.	.	4	.	.	.	.	.	.	.	3
346	7401030-A001/001		.	.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	1
347	7401032-A001/001		2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
348	7401066-A001/001		.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	1
349	7401067-A001/001	N003001	.	.	.	.	.	.	.	.	12	.	.	13	.	.	.	.	.	15	.	3
350	7401067-A002/002	N003002	.	.	.	.	.	.	.	.	12	.	.	13	.	.	.	.	.	16	.	3
351	7401067-A003/003	N003003	.	.	.	.	.	.	.	.	12	.	.	13	.	.	.	.	.	17	.	3
353	7401069-A001/001		.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	1
357	7401073-A001/001		.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
358	7401074-A001/001	N001401	.	.	.	.	.	3	.	.	.	11	.	.	12	.	.	.	.	21	.	4
359	7401075-A001/001		.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	1
362	7401078-A001/001		.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1

Table B-6

IRT TREND ANALYSIS ITEM TABLE FOR AGE 13

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2 1	2 2	2 3	2 4	2 5	2 11	2 12	2 13	6 1	6 2	6 3	11 1	11 2	11 3	11 14	15 1	15 2	15 3	15 4	NO. YEARS USED
364	7401080-A001/001							10														1
367	7401083-A001/001		12																			1
371	7402020-A001/001					1					6			7								3
375	7403007-A001/001	N004901					14					15			16					18		4
377	7403019-A001/001	N002501									14			15					17			3
380	7503004-A001/001					9																1
385	7503044-A001/001				16								12									2
386	7503045-A001/001		7										3			3						3
404	H222000-A001/001	N001601															8			8		2
405	H222000-A002/002	N001602															8			9		2
406	H222000-A003/003																8					1
407	H222000-A004/004																8					1
408	H222000-A005/005	N001604															8			11		2
417	H241000-A001/001	N004401															3				15	2
418	H241000-A002/002																3					1
419	H241000-A003/003																3					1
433	H265000-A001/001	N002001																	11			1
434	H265000-A002/002	N002002																	12			1
444	H284000-A001/001	N003201																7				1
447	H284000-A004/004	N003204																10				1
448	H286000-A001/001	N004701																15				1
449	H286000-A002/002	N004702																16				1
450	H286000-A003/003	N004703																17				1
463	H403000-A001/001	N007301															6	1				2
464	H403000-A002/002	N007302															6	2				2
465	H403000-A003/003	N007303															6	3				2
466	H403000-A004/004	N007304															6	4				2
467	H403000-A005/005	N007305															6	5				2
468	H403000-A006/006	N007306															6	6				2
495	H413000-A002/002	N008202																		2		1
496	H413000-A003/003	N008203																		3		1
497	H413000-A004/004	N008204																		4		1
634		N003202																8				1
635		N003203																9				1
636		N005101																	2			1
637		N005001																	6			1
638		N001702																	4			1
639		N001703																	5			1
640		N002003																	13			1
641		N003301																		19		1
642		N001202																		27		1
643		N004801																		20		1
644		N003901																		14		1
645		N008201																		1		1
646		N008205																		5		1
647		N001603																		10		1
648		N003802																			3	1
649		N003803																			4	1
650		N004201																			18	1
651		N004202																			19	1

Table B-6

IRT TREND ANALYSIS ITEM TABLE FOR AGE 13

POS	ECS ID	\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	1	2	3	4	5	11	12	13	1	2	3	1	2	3	14	1	2	3	4	YEARS
		NAEP ID\																				USED
652		N002102	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	1
654		N004402	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	1
655		N004403	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	1
NUMBER OF CALIBRATED ITEMS IN BOOKLET			19	15	15	15	9	20	34	15	30	19	30	26	15	22	14	15	15	19	14	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET ACROSS YEARS			6	10	9	6	2	8	10	5	29	19	30	26	15	22	10	8	8	9	7	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET WITHIN YEARS			0	0	0	0	0	0	0	0	4	4	4	4	4	4	0	0	0	0	0	

Table B-7

IRT TREND ANALYSIS ITEM TABLE FOR AGE 17

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	2	3	4	5	7	8	9	10	1	2	3	1	2	3	11	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																				USED
5	7099003-A001/001			15																		1
6	7099004-A001/001										2			2								2
10	7099006-A001/001		5								6			8								3
11	7099006-A002/002		5								6			8								3
12	7099007-A001/001						13						6			6						3
13	7099007-A002/002	N005002					13						6			6			7			4
14	7099007-A003/003	N005003					13						6			6			8			4
16	7099009-A001/001	N003601				7						2			2			13				4
17	7099009-A002/002	N003602				7						2			2			14				4
19	7099012-A001/001	N003501	1								11			14					10			4
20	7099013-A001/001						9															1
21	7099013-A002/002						9															1
22	7099013-A003/003						9															1
48	7099020-A001/001								2													1
49	7099020-A002/002								2													1
50	7099020-A003/003								2													1
52	7099021-A002/002							11														1
53	7099021-A003/003							11														1
54	7099021-A004/004							11														1
57	7099023-A001/001									9												1
58	7099023-A002/002									9												1
59	7099023-A003/003									9												1
62	7099026-A001/001			6							4			5								3
63	7099026-A002/002			6							4											2
64	7099026-A003/003			6							4											2
65	7099027-A001/001				7							7			7							3
67	7101009-A001/001								1				8									2
69	7101019-A001/002					18																1
70	7101019-A003/004					18																1
71	7101019-A005/006					18																1
72	7101019-A007/008					18																1
73	7101019-A009/010					18																1
74	7101019-A011/012					18																1
75	7101019-A013/014					18																1
94	7102013-A001/001										19	20	18	20	20	18						2
95	7102014-A001/001							6														1
96	7102015-A001/001					1																1
107	7102035-A001/001			12									9			9						3
108	7102036-A001/001					3																1
109	7102037-A001/001							12														1
110	7102037-A002/002							12														1
113	7103004-A001/001							2														1
114	7103012-A001/001						11															1
115	7103017-A001/001									6												1
124	7103030-A001/001							3														1
125	7103031-A001/001						15															1
126	7103032-A001/001					8						1			1							3
133	7103039-A001/001											9			9							2
134	7103041-A001/001									11			7			7						3

Table B-7

IRT TREND ANALYSIS ITEM TABLE FOR AGE 17

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2 2	2 3	2 4	2 5	2 7	2 8	2 9	2 10	6 1	6 2	6 3	11 1	11 2	11 3	11 11	15 1	15 2	15 3	15 4	NO. YEARS USED
135	7103041-A002/002									11			7			7						3
136	7103041-A003/003									11			7			7						3
137	7103041-A004/004									11			7			7						3
138	7103042-A001/001					12																1
139	7103042-A002/002					12																1
140	7103043-A001/001						10															1
141	7103044-A001/001	N001701							9				5			5			3			4
142	7103044-A002/002								9				5			5						3
143	7103044-A003/003								9				5			5						3
144	7103045-A001/001	N005201				5						10			10						23	4
146	7103045-A003/003	N005203				5						10			10						25	4
147	7103046-A001/001						4															1
148	7103046-A002/002						4															1
151	7103049-A001/001		13																			1
152	7103049-A002/002		13																			1
153	7103050-A001/001				16							8			8							3
156	7103052-A001/001						14						4			4						3
162	7103058-A001/001		11																			1
163	7103059-A001/001						5															1
164	7103060-A001/001									1												1
166	7103062-A001/001		10																			1
167	7103062-A002/002		10																			1
168	7103062-A003/003		10																			1
169	7127001-A001/001	N003801									9			12							2	3
170	7127001-A002/002										9			12								2
171	7127001-A003/003										9			12								2
174	7127003-A001/001	N002101								2			10			10					5	4
176	7127004-A001/001				12																	1
177	7127004-A002/002				12																	1
180	7127006-A001/001		3																			1
181	7127007-A001/001			8							20	21	19	21	21	19						3
182	7127009-A001/002							15					15									2
183	7127009-A003/004							15					15									2
184	7127009-A005/005							15					15									2
185	7127009-A006/006							15					15									2
186	7127009-A007/007							15					15									2
187	7127009-A008/008							15					15									2
188	7127009-A009/009							15					15									2
191	7201013-A002/003					10																1
193	7201023-A001/001									5												1
194	7201024-A001/001								7													1
195	7201025-A001/001				10																	1
196	7202003-A001/001										3			3								2
197	7202008-A001/001		14								18	19	17	19	19	17						3
200	7203006-A001/001				11																	1
201	7203009-A001/001			4							12			15								3
202	7203009AA001/001							13														1
203	7203010-A001/001			7																		1
204	7203011-A001/001						17															1
205	7203012-A001/001												11			12						2

Table E-7

IRT TREND ANALYSIS ITEM TABLE FOR AGE 17

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO. YEARS USED
			2	3	4	5	7	8	9	10	1	2	3	1	2	3	11	1	2	3	4	
206	7203013-A001/001		.	.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	1
210	7203046-A001/001		.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
212	7203048-A001/001		.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
213	7203049-A001/001		.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	1
216	7203052-A001/001		.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	1
217	7203053-A001/001		.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
236	7301022-A001/002		.	.	.	.	.	.	12	.	15	.	.	.	.	.	.	.	.	.	.	2
237	7301022-A003/004		.	.	.	.	.	.	12	.	15	.	.	.	.	.	.	.	.	.	.	2
238	7301022-A005/006		.	.	.	.	.	.	12	.	15	.	.	.	.	.	.	.	.	.	.	2
239	7301022-A009/010		.	.	.	.	.	.	12	.	15	.	.	.	.	.	.	.	.	.	.	2
240	7301027-A001/002		15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
241	7301027-A003/004		15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
242	7301027-A005/006		15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
254	7302008-A001/002		.	.	.	.	18	.	.	.	.	.	13	.	.	15	.	.	.	.	.	3
255	7302009-A001/002		.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	1
256	7302009-A003/004		.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	1
257	7302009-A005/006		.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	1
258	7302009-A007/008		.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	1
259	7302009-A009/010		.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	1
281	7303017-A001/001		.	14	.	.	.	.	.	.	10	.	.	13	.	.	.	.	.	.	.	3
282	7303017-A002/002		.	14	.	.	.	.	.	.	10	.	.	13	.	.	.	.	.	.	.	3
288	7303019-A001/001	N001201	.	.	.	.	8	.	.	.	.	.	2	.	.	2	.	.	.	26	.	4
289	7303019-A002/002		.	.	.	.	8	.	.	.	.	.	2	.	.	2	.	.	.	.	.	3
293	7303026-A003/004		.	.	.	.	.	.	15	.	16	.	.	.	.	.	.	.	.	.	.	2
294	7303026-A005/005		.	.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	1
310	7303051-A001/001	N002201	.	2	.	.	.	.	.	.	8	.	.	10	.	.	.	.	14	.	.	4
311	7303051-A002/002	N002202	.	2	.	.	.	.	.	.	8	.	.	10	.	.	.	.	15	.	.	4
312	7303051-A003/003	N002203	.	2	.	.	.	.	.	.	8	.	.	10	.	.	.	.	16	.	.	4
314	7303054-A001/001		.	.	.	13	.	.	.	.	.	13	.	.	13	.	.	.	.	.	.	3
315	7303054-A002/002		.	.	.	13	.	.	.	.	.	13	.	.	13	.	.	.	.	.	.	3
316	7303055-A002/002		.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
317	7303055-A003/003		.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
318	7303056-A001/002		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	1
319	7303056-A003/004		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	1
320	7303056-A005/006		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	1
321	7303056-A007/008		.	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	1
322	7303057-A001/002		.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
323	7303057-A003/004		.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
324	7303057-A005/006		.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
325	7303057-A007/008		.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
339	7401007-A001/001		.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
345	7401024-A001/001		.	.	.	9	.	.	.	.	.	11	.	.	11	.	.	.	.	.	.	3
346	7401030-A001/001		.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
347	7401032-A001/001		.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
349	7401067-A001/001	N003001	6	.	.	.	.	.	.	.	7	.	.	9	.	.	.	.	.	15	.	4
350	7401067-A002/002	N003002	6	.	.	.	.	.	.	.	7	.	.	9	.	.	.	.	.	16	.	4
351	7401067-A003/003	N003003	6	.	.	.	.	.	.	.	7	.	.	9	.	.	.	.	.	17	.	4
354	7401070-A001/001		.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
358	7401074-A001/001	N001401	.	.	1	.	.	.	.	.	.	3	.	.	3	.	.	.	.	21	.	4
360	7401076-A001/001		8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Table B-7

IRT TREND ANALYSIS ITEM TABLE FOR AGE 17

POS	ECS ID	\ YEAR \PACKAGE NAEP ID\	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO. YEARS USED
			2	3	4	5	7	8	9	10	1	2	3	1	2	3	11	1	2	3	4	
363	7401079-A001/001				14																	1
364	7401080-A001/001								3				3			3						3
367	7401083-A001/001				2							12			12							3
375	7403007-A001/001	N004901				14						5			5					18		4
377	7403019-A001/001	N002501						4					12			13			17			4
381	7503005-A001/001									4												1
385	7503044-A001/001				3							6										2
386	7503045-A001/001										1			1								2
390	H202000-A001/001																4					1
391	H202000-A002/002																4					1
392	H202000-A003/003																4					1
393	H202000-A004/004																4					1
433	H265000-A001/001	N002001																	11			1
434	H265000-A002/002	N002002																	12			1
444	H284000-A001/001	N003201																7				1
447	H284000-A004/004	N003204																10				1
463	H403000-A001/001	N007301															6	1				2
464	H403000-A002/002	N007302															6	2				2
465	H403000-A003/003	N007303															6	3				2
466	H403000-A004/004	N007304															6	4				2
467	H403000-A005/005	N007305															6	5				2
468	H403000-A006/006	N007306															6	6				2
495	H413000-A002/002	N008202																		2		1
496	H413000-A003/003	N008203																		3		1
497	H413000-A004/004	N008204																		4		1
516	H441000-A001/001																7					1
517	H441000-A002/002																7					1
518	H441000-A003/003																7					1
634		N003202																8				1
635		N003203																9				1
636		N005101																	2			1
637		N005001																	6			1
638		N001702																	4			1
639		N001703																	5			1
640		N002003																	13			1
641		N003301																		19		1
642		N001202																		27		1
643		N004801																		20		1
644		N003901																		14		1
645		N008201																		1		1
646		N008205																		5		1
647		N001603																		10		1
648		N003802																			3	1
649		N003803																			4	1
650		N004201																			18	1
651		N004202																			19	1
652		N002102																			6	1
653		N008108																			14	1

Table B-7

IRT TREND ANALYSIS ITEM TABLE FOR AGE 17

		\ YEAR	2	2	2	2	2	2	2	2	6	6	6	11	11	11	11	15	15	15	15	NO.
		\PACKAGE	2	3	4	5	7	8	9	10	1	2	3	1	2	3	11	1	2	3	4	YEARS
POS	ECS ID	NAEP ID\																				USED
654		N004402	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	1
655		N004403	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	1
NUMBER OF CALIBRATED ITEMS IN BOOKLET			18	21	10	22	21	23	14	18	30	18	30	22	17	22	13	12	15	16	12	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET ACROSS YEARS			7	11	5	9	7	8	10	5	29	18	30	22	17	22	6	8	8	6	4	
NUMBER OF CALIBRATED ITEMS LINKING BOOKLET WITHIN YEARS			0	0	0	0	0	0	0	0	3	3	3	3	3	3	0	0	0	0	0	

Table B-8

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
1	N001101	0.34386	0.04756	-0.38421	0.10402	0.29068	0.05300
2	N001201	0.71152	0.18319	1.14426	0.40276	0.36925	0.05312
3	N001202	1.27637	0.18712	0.58486	0.19679	0.25554	0.03702
4	N001301	0.98591	0.11808	0.49490	0.17315	0.40040	0.04266
5	N001302	0.71972	0.08748	-1.54817	0.24087	0.49736	0.09718
6	N001303	1.53406	0.13290	0.40675	0.13115	0.28084	0.03014
7	N001401	0.99901	0.09413	0.00071	0.11426	0.25110	0.05479
8	N001501	1.80824	0.13010	-1.31319	0.15232	0.22542	0.04742
9	N001502	1.64315	0.09754	-0.50728	0.06080	0.18160	0.02631
10	N001503	1.34461	0.08765	-0.90181	0.08628	0.20744	0.04272
11	N001504	1.44778	0.08909	-0.65029	0.06788	0.17293	0.03234
12	N001506	0.65437	0.04285	2.07853	0.14371	0.00000	0.00000
13	N001601	0.62223	0.04094	-0.95892	0.08244	0.13299	0.04576
14	N001602	1.26259	0.07887	-0.69250	0.06605	0.25043	0.03097
15	N001603	0.81562	0.07272	-0.03084	0.06802	0.23261	0.03311
16	N001604	1.37491	0.10123	0.11143	0.06294	0.26918	0.01844
17	N001701	0.98126	0.06647	-0.41778	0.09050	0.23118	0.05878
18	N001702	0.54099	0.11563	2.65058	0.62070	0.23141	0.02826
19	N001703	1.08107	0.08003	0.00329	0.09642	0.29079	0.04417
21	N001802	1.59211	0.14013	0.72655	0.13110	0.21661	0.01183
22	N001901	1.64357	0.11147	0.20973	0.09766	0.33071	0.02786
23	N001903	0.93422	0.03812	0.15576	0.02868	0.00000	0.00000
24	N002001	1.19684	0.06511	-0.01271	0.04963	0.13080	0.02011
25	N002002	1.44381	0.08389	-0.04179	0.05539	0.20290	0.02023
26	N002003	1.58260	0.09278	-0.22910	0.05407	0.22409	0.02191
27	N002101	0.94082	0.09421	1.17114	0.17573	0.24676	0.01860
28	N002102	1.49484	0.09999	0.84044	0.11809	0.14741	0.01242
29	N002201	1.70361	0.11834	-0.12913	0.07771	0.20050	0.03659
30	N002202	1.35786	0.11962	-0.34881	0.11154	0.33683	0.05918
31	N002203	0.78303	0.06634	-1.13922	0.13696	0.23625	0.08550
32	N002401	1.44902	0.09561	-0.37505	0.05657	0.12753	0.02317
33	N002501	0.54969	0.05253	0.12918	0.10047	0.20489	0.05712
34	N002701	1.02419	0.10249	0.83348	0.16436	0.23428	0.03194
35	N002702	1.14818	0.07671	0.05508	0.06496	0.14060	0.02286
36	N002801	1.92053	0.11379	-0.76744	0.08091	0.17456	0.02795
37	N002802	1.89576	0.10954	-0.91192	0.09238	0.14345	0.02805
38	N002803	0.33105	0.02801	2.20046	0.18826	0.00000	0.00000
39	N002902	0.55751	0.04967	-0.80149	0.11433	0.22919	0.07069
40	N002903	2.31343	0.18003	-0.34122	0.08178	0.25293	0.03950
41	N002904	1.28934	0.09451	-0.02029	0.08728	0.19746	0.04124
42	N002905	0.75794	0.05783	0.24793	0.08254	0.11561	0.04032
43	N002906	1.96425	0.14802	-0.36322	0.08228	0.23041	0.04409
44	N003001	1.29316	0.10896	1.15281	0.16887	0.20713	0.01285

Table B-8
(continued)

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
45	N003002	0.30916	0.02912	0.11920	0.06518	0.16796	0.04095
46	N003003	2.29397	0.10854	1.72388	0.18958	0.11968	0.00632
47	N003101	1.57066	0.10038	-0.64538	0.07260	0.26709	0.03160
48	N003102	1.53025	0.08338	-0.35908	0.05122	0.14507	0.02252
49	N003103	0.70382	0.04221	1.92324	0.12434	0.00000	0.00000
50	N003201	1.20709	0.08792	-0.59316	0.08731	0.17078	0.05590
51	N003202	1.59047	0.12373	0.01181	0.09272	0.22727	0.03782
52	N003203	1.21513	0.10141	0.23965	0.10682	0.22248	0.03861
53	N003204	1.45667	0.11979	0.25952	0.11171	0.23829	0.03463
54	N003301	1.14150	0.08108	-0.40955	0.07830	0.15767	0.04884
55	N003401	1.46724	0.14986	-0.20675	0.09152	0.15919	0.04704
56	N003501	0.75144	0.06201	-0.44828	0.09301	0.17212	0.06140
57	N003601	1.45231	0.11616	-0.66817	0.09937	0.20310	0.05986
58	N003602	1.31985	0.10929	-0.13019	0.09684	0.24110	0.04772
59	N003701	0.73641	0.06121	-0.76037	0.10437	0.23915	0.05983
60	N003702	1.07106	0.08364	-0.01031	0.07761	0.23644	0.03157
61	N003703	0.68872	0.03648	0.29437	0.03458	0.00000	0.00000
62	N003801	0.89130	0.11188	1.46486	0.25097	0.30911	0.01789
63	N003802	0.41376	0.02983	-0.70293	0.07824	0.11010	0.04749
64	N003803	0.75737	0.09309	1.60016	0.24505	0.20621	0.01882
65	N003901	1.37453	0.19223	-1.84669	0.33113	0.23174	0.08938
66	N004002	0.61451	0.07927	-1.42643	0.21398	0.24613	0.09327
67	N004101	1.09618	0.08674	-1.12198	0.11390	0.22894	0.05358
68	N004201	1.10307	0.07055	0.03059	0.06151	0.18470	0.02437
69	N004202	0.76199	0.07167	0.18683	0.09769	0.29069	0.03786
70	N004301	1.41953	0.12483	0.40405	0.13117	0.28769	0.03220
71	N004302	0.62069	0.04330	0.58170	0.05603	0.00000	0.00000
72	N004401	1.71824	0.12695	-1.77446	0.20185	0.26243	0.06523
73	N004402	0.87572	0.07482	-0.22006	0.06568	0.14810	0.03611
74	N004403	1.64193	0.12805	-1.46711	0.17026	0.22811	0.05425
75	N004501	0.97362	0.10339	0.49345	0.15138	0.30495	0.04322
76	N004502	0.68013	0.05431	-0.82441	0.10455	0.17982	0.06770
77	N004601	0.89929	0.07817	0.17933	0.10423	0.18399	0.04832
78	N004602	1.31823	0.10310	-0.08468	0.09194	0.24890	0.04420
79	N004603	1.48506	0.11293	-0.51585	0.08941	0.22557	0.05438
80	N004604	0.79691	0.04948	-0.61689	0.05145	0.00000	0.00000
81	N004701	1.69375	0.10145	-0.51490	0.05860	0.20392	0.02105
82	N004702	0.76376	0.06512	-0.92837	0.10528	0.23709	0.05750
83	N004703	1.02059	0.06467	-0.65101	0.06202	0.15261	0.03124
84	N004801	1.25733	0.08452	-1.25766	0.10848	0.24193	0.04726
85	N004901	0.91600	0.05695	0.22127	0.05969	0.19010	0.02127
86	N005001	1.99291	0.10250	1.37994	0.15923	0.21076	0.01059

Table B-8
(continued)

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
87	N005002	0.85936	0.10848	1.28831	0.24040	0.26406	0.02938
88	N005003	0.73710	0.10541	1.90503	0.33064	0.13504	0.02371
89	N005101	0.84190	0.06067	-2.13982	0.17848	0.23555	0.08295
90	N005201	0.67359	0.10746	0.63625	0.22965	0.48066	0.05423
91	N005202	0.59985	0.07072	0.58164	0.15242	0.20628	0.05831
92	N005203	1.14306	0.12119	1.83713	0.28411	0.30875	0.01452
93	N005301	1.13281	0.14635	-0.02794	0.13187	0.28349	0.05862
94	N005302	1.40569	0.14546	0.38677	0.11915	0.12855	0.02959
95	N005303	0.86748	0.19494	1.00844	0.34366	0.32993	0.04769
96	N005304	1.81002	0.19697	0.05192	0.11407	0.22677	0.03838
97	N005305	1.08554	0.12148	-0.67673	0.13021	0.22225	0.07653
98	N005403	1.34660	0.15280	-0.33453	0.11463	0.28866	0.06065
99	N005404	1.45537	0.13848	-1.03748	0.14409	0.18712	0.06678
100	N005405	2.01849	0.19497	0.06798	0.09988	0.20587	0.03105
101	N005406	1.20953	0.11578	-0.39768	0.09352	0.18463	0.05446
102	N005407	1.77747	0.20114	-0.24601	0.11000	0.32637	0.04931
103	N005503	0.71843	0.07420	0.35569	0.12684	0.21105	0.05387
104	N005504	1.31644	0.11181	0.77755	0.14729	0.21947	0.02374
105	N005505	1.12595	0.09159	-0.91282	0.12097	0.24680	0.07913
106	N005601	1.38681	0.15119	-0.65281	0.12495	0.25274	0.07133
107	N005602	1.71547	0.18749	0.29673	0.13264	0.20783	0.03130
108	N005603	1.48703	0.17122	-0.17746	0.11343	0.30619	0.05075
109	N007301	1.18343	0.09087	-0.39389	0.09968	0.27841	0.05857
110	N007302	0.81787	0.05868	0.28535	0.08412	0.13646	0.03885
111	N007303	1.10993	0.07680	-0.02429	0.08404	0.19644	0.04296
112	N007304	0.88667	0.07155	-0.00694	0.09984	0.22304	0.05305
113	N007305	0.52937	0.04195	0.01012	0.07697	0.13321	0.04958
114	N007306	1.00946	0.05679	-0.11609	0.05916	0.10318	0.03478
115	N007401	1.09780	0.07624	0.53070	0.09581	0.12305	0.02729
116	N007402	1.30369	0.08423	-0.31735	0.07522	0.17614	0.04544
117	N007403	1.75588	0.11861	0.21398	0.09309	0.23342	0.02691
118	N007404	0.98461	0.07227	0.05998	0.08757	0.18069	0.04382
119	N007405	0.88671	0.10164	1.40133	0.22916	0.18663	0.02508
120	N007407	0.85127	0.04249	0.67125	0.04940	0.00000	0.00000
135	N008201	2.72430	0.30170	-0.47105	0.13058	0.32310	0.05201
136	N008202	1.14590	0.10210	-0.06540	0.10238	0.18767	0.05219
137	N008203	1.54336	0.14080	-0.28855	0.10359	0.24721	0.05374
138	N008204	2.59971	0.23614	-0.22772	0.09180	0.20928	0.03761
139	N008205	2.14522	0.18750	-0.25591	0.09158	0.20457	0.04198
140	N008207	0.59778	0.06018	2.25854	0.23732	0.00000	0.00000
141	N008601	1.78949	0.17931	-0.97240	0.17067	0.16947	0.03727
142	N008602	1.36797	0.17927	-0.55448	0.12210	0.26122	0.04075

Table B-8
(continued)

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
143	N008603	1.20570	0.11778	-0.98525	0.13713	0.14020	0.04300
144	N008701	1.19247	0.13395	-2.39050	0.34204	0.24019	0.08756
145	N008801	1.48897	0.10032	-1.78932	0.17324	0.19373	0.05559
146	N008901	1.32842	0.10576	-1.24354	0.13802	0.14826	0.04118
147	N008902	1.25848	0.10203	-1.27067	0.13970	0.15626	0.04339
148	N008904	0.67255	0.06421	-2.50923	0.25598	0.00000	0.00000
149	N009001	1.32810	0.15196	-0.43338	0.09699	0.15397	0.03091
150	N009002	1.17681	0.16319	-0.09320	0.08727	0.17778	0.02989
151	N009003	0.84446	0.20302	0.75488	0.24155	0.22631	0.03200
152	N009004	1.76785	0.22483	-0.34990	0.10912	0.23959	0.02739
153	N009101	1.00715	0.12010	-1.45087	0.21000	0.25570	0.07574
154	N009201	1.79466	0.17188	-1.37708	0.21643	0.30102	0.05431
155	N009401	1.88222	0.12694	-1.40233	0.17247	0.10485	0.03564
156	N009601	1.36014	0.10602	-1.87239	0.20702	0.12966	0.05296
157	N009701	1.08207	0.12379	-0.65381	0.11174	0.16395	0.04143
158	N009702	1.95945	0.22710	-0.53268	0.13089	0.24930	0.02757
159	N009703	1.44885	0.21085	-0.16529	0.09722	0.25808	0.02879
160	N009704	1.14952	0.18527	0.03287	0.09564	0.20890	0.03109
161	N009705	1.95691	0.20749	-0.70162	0.14664	0.21082	0.02946
162	N009801	1.39630	0.13433	-2.22709	0.29571	0.25930	0.08622
163	N009901	0.97641	0.11673	-1.04928	0.15973	0.20639	0.05932
164	N010002	1.29007	0.13727	-1.09449	0.16480	0.17234	0.04708
165	N010003	1.65704	0.19421	-0.94025	0.17915	0.24113	0.04220
166	N010102	1.12440	0.19308	-0.04987	0.11066	0.26735	0.03663
167	N010103	1.79514	0.20006	-1.07518	0.20659	0.20880	0.04202
168	N010201	1.24259	0.12137	-1.93207	0.24523	0.24403	0.07778
169	N010301	0.70206	0.08494	-2.38310	0.31786	0.24783	0.09312
170	N010401	0.71532	0.08678	-1.48668	0.20893	0.21889	0.07731
171	N010402	0.92807	0.17080	0.13218	0.11255	0.22214	0.03717
172	N010403	1.03079	0.19719	0.46460	0.15326	0.18965	0.02710
173	N010501	2.02312	0.13916	-1.49019	0.19000	0.20367	0.04622
174	N010502	1.20386	0.11435	-1.19606	0.15414	0.15642	0.04854
175	N010503	1.45500	0.12343	-1.45957	0.18367	0.15940	0.04832
176	N010504	2.30030	0.16567	-1.11438	0.17388	0.18153	0.03226
177	N010601	1.60441	0.19642	-0.63370	0.13572	0.24626	0.03552
178	N010602	1.78849	0.34380	0.20854	0.15336	0.30609	0.02271
179	N010603	1.35893	0.19872	-0.25829	0.10519	0.23369	0.03268
180	N010604	1.63695	0.25020	-0.10064	0.10586	0.23522	0.02634
181	N010605	1.21990	0.19002	-0.06859	0.09829	0.18417	0.03126
182	N010701	1.06419	0.12219	-1.28254	0.13812	0.17860	0.06280
183	N010801	1.08381	0.11919	-0.47131	0.08698	0.26010	0.03530
184	N010902	1.56357	0.15281	-0.46713	0.08701	0.24105	0.02630

Table B-8
(continued)

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
185	N010903	1.85005	0.15682	-0.56355	0.09617	0.19270	0.02240
186	N010904	1.52152	0.17022	-0.24453	0.08018	0.27499	0.02425
187	N011001	1.27933	0.11570	-0.87939	0.11252	0.22839	0.03695
188	N011002	1.65730	0.16726	-0.31498	0.08113	0.25211	0.02193
189	N011003	2.41596	0.16943	-0.92768	0.14716	0.24143	0.02440
190	N011004	1.78824	0.15874	-0.54263	0.09661	0.22609	0.02276
191	N011101	1.56839	0.14127	-0.54055	0.08967	0.19668	0.02517
192	N011201	0.91063	0.11742	-0.25916	0.08529	0.25976	0.03693
193	N011301	1.65295	0.14312	-0.75604	0.11135	0.21079	0.02826
194	N011302	0.99244	0.11855	-0.42973	0.08894	0.22689	0.03887
195	N011401	0.83827	0.19405	0.69656	0.22734	0.33376	0.02999
196	N011402	0.82218	0.13917	0.01027	0.10228	0.28837	0.04148
197	N011403	0.97146	0.19233	0.62100	0.18629	0.26972	0.02544
198	N011404	1.32661	0.21546	0.49246	0.15146	0.21998	0.01852
199	N012901	1.09100	0.09082	-1.60092	0.16886	0.13573	0.05331
200	N013001	1.01980	0.12202	-0.34301	0.08344	0.16456	0.03631
201	N013002	0.97225	0.12117	-0.38259	0.08966	0.18704	0.03960
202	N013003	1.71689	0.16408	-1.12308	0.17239	0.23372	0.04150
203	N013004	0.99397	0.11477	-0.94626	0.13780	0.21625	0.05576
204	N013101	1.75676	0.13641	-1.56003	0.19617	0.21484	0.05240
205	N013102	1.40116	0.14636	-0.78899	0.12368	0.21834	0.03905
206	N013103	0.95398	0.09740	-0.86824	0.11733	0.14741	0.04563
207	N013104	0.75969	0.11642	-0.42068	0.11125	0.21563	0.05536
208	N013201	1.66526	0.21104	-0.69299	0.16001	0.18120	0.03670
209	N013301	1.23192	0.16119	-1.55713	0.26807	0.25257	0.07686
210	N013401	1.20290	0.17717	-0.25020	0.10663	0.15710	0.03510
211	N013402	1.43756	0.18853	-0.86245	0.17513	0.20481	0.04755
212	N013403	1.49438	0.22338	-0.27786	0.11622	0.19858	0.03309
222	N014001	1.23795	0.15297	-0.85718	0.14869	0.24936	0.04816
223	N014101	0.75814	0.07084	-1.28284	0.14222	0.16926	0.05891
224	N014201	1.20695	0.13433	-1.21830	0.18886	0.13647	0.05202
225	N014301	1.75466	0.19139	-0.81991	0.15771	0.19017	0.03482
226	N014302	1.07393	0.13609	-0.49827	0.10782	0.18111	0.04062
227	N014303	1.72083	0.18651	-1.04120	0.18841	0.20824	0.04131
228	N014501	0.43233	0.06452	-2.26356	0.34802	0.00000	0.00000
229	N014502	0.93414	0.12331	-2.66383	0.40566	0.00000	0.00000
230	N014503	0.62434	0.13263	-4.12019	0.90265	0.00000	0.00000
231	N015101	0.93155	0.11001	0.34271	0.16850	0.23395	0.06686
232	N015102	2.53320	0.23579	0.54834	0.20683	0.21641	0.03021
233	N015103	2.40149	0.19982	0.66008	0.19694	0.21907	0.02802
234	N015104	1.70738	0.18424	0.44096	0.19311	0.27788	0.04502
235	N015201	1.08863	0.12589	-0.76561	0.14968	0.22684	0.08504

Table B-8
(continued)

Item Parameter Estimates and Standard Errors

Item	ETS ID	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
236	N015502	1.27279	0.12588	0.18925	0.14019	0.20864	0.05669
237	N015503	0.91211	0.11941	0.75605	0.21582	0.24651	0.05565
238	N015504	1.18882	0.12068	0.10997	0.13819	0.22004	0.06172
239	N015505	0.68340	0.08254	-0.17492	0.14577	0.24726	0.08705
240	N015901	1.02091	0.13331	0.37058	0.20437	0.33263	0.06812
241	N015902	1.38041	0.16520	0.72633	0.23435	0.31666	0.04295
242	N015903	1.18177	0.12895	1.10148	0.22398	0.15255	0.03180
243	N015904	0.65706	0.06232	0.85809	0.10582	0.00000	0.00000
244	N016001	1.04259	0.12162	0.03342	0.16386	0.28516	0.07766
245	N016002	1.38569	0.15434	1.24735	0.27581	0.45609	0.02768
246	N016003	0.90591	0.10257	0.35438	0.15749	0.20510	0.06497
247	N016004	1.09517	0.12568	0.10257	0.16454	0.27136	0.07392
248	N016005	1.73407	0.17470	0.15598	0.15207	0.23040	0.05394
249	N016006	1.35727	0.13670	0.42447	0.16122	0.20306	0.04911
250	N017001	1.51834	0.15713	0.48407	0.17457	0.21320	0.04241
251	N017002	1.93512	0.13762	1.10006	0.19253	0.19574	0.02171
252	N017003	1.83349	0.12901	1.76996	0.24850	0.17677	0.01566

GLOSSARY

GLOSSARY

administration. The conduct of a National Assessment session.

Administration Schedule. A list of the name, age and sex of each student invited to a particular assessment session.

administration time. The total time allowed for an item. (Includes the time allowed for the stimulus and the response.)

administration timetable. Time periods during the school year when the various grade/age groups are assessed. The time periods for the Year 15 assessment were:

Grade 8/Age 13
October 10 to December 16, 1983

Grade 4/Age 9
January 2 to March 9, 1984

Grade 11/Age 17
March 12 to May 11, 1984

administrative units. Geographic areas such as states, counties, school districts, etc.

AERA. American Educational Research Association.

age-eligible. An individual who meets the age definition for one of the National Assessment populations: 9-year-olds, 13-year-olds, 17-year-olds.

aggregate estimate. Estimate for a combination of smaller groups for which estimates have been produced.

allocation. Apportionment of a total sample size to various parts of the population (See *final allocation.*)

almanacs. The sets of tables summarizing NAEP results.

anchoring. The process of characterizing score levels in terms of predicted observable behavior.

ARM. See *Average Response Method.*

assessment. The documentation of the progress in knowledge, skills and attitudes of American youth. Measures are taken at periodic intervals for each learning area, with the goal of determining trends and reporting the findings to the public and to the education community. See also *National Assessment of Educational Progress.*

assessment administrator. Individual employed to administer the assessment in participating schools.

assessment session. The period of time during which a NAEP package is administered to one or more individuals.

Average Response Method. A regress-based technique for predicting for a respondent the conditional distribution of an average score on a set of exercises given responses to at least one of the exercises and other information. Used to produce the NAEP Year 15 Writing Scale.

average sample size. The average sample obtained per sampling unit selected.

background and attitude items. See *non-cognitive assessment.*

bias. In statistics, the difference between the expected value of an estimator and the population parameter being estimated. If the average value of the estimator over all possible samples (the estimator's expected value) equals the parameter being estimated, the estimator is said to be **unbiased**; otherwise, the estimator is **biased**.

BIB (Balanced Incomplete Block) spiralling. A complex variant of multiple matrix sampling, in which a small subset of items is administered to each respondent in such a way that each pair of items is administered to a nationally representative subsample of respondents.

BILOG. A computer program for estimating item parameters by marginal estimation procedures.

block. A group of assessment items created by dividing the item pool for a grade/age into subsets. Used in the implementation of the BIB and UBIB Spiral sample design.

booklet. The assessment instrument created by combining blocks of assessment items.

bridging. An administration of the same set of exercises under two different conditions or to two different populations to allow a statistical link (*"bridge"*) to be established between results under the different circumstances.

calibrate. To estimate the parameters of a set of items from responses of a sample of a set of examinees.

category (scoring). A classification of a response to an open-ended item. See *Scoring Guide.*

category within a variable. A sub-classification within a variable, or subgroup. For example, Male and Female are categories of the subgroup Sex. See *Reporting Subgroups.*

cell. The smallest unit of a table. For example, a two-way table with 5 rows and 7 columns contains 35 cells (5 x 7 = 35).

census tract (CT). Small, relatively permanent areas into which large cities and adjacent areas are divided for the purpose of providing small-area statistics. The average census tract contains approximately 4,000 residents.

clustering. The process of forming sampling units as groups of other units.

codebook. A printout of the raw data files for each student, excluded student, teacher and school in a particular grade/age.

coefficient of variation. The ratio of the standard deviation of an estimate to the value of the estimate.

cognitive assessment. The portion of the Year 15 NAEP which assessed students' abilities in the learning areas of reading and writing.

combined ratio estimator. The ratio estimator resulting from first estimating the numerator and the denominator values and then using the quotient of these as the estimate of the ratio.

common block. A group of background items included in the beginning of every assessment booklet.

complete enumeration survey. Survey in which the entire population is enumerated or surveyed; a census.

conditional probability. Probability of an event, given the occurrence of another event.

conditioning variables. Demographic variables characterizing respondent. Used in construction of plausible values.

controlled selection. A method of probability sampling involving balanced samples on asymmetrical controls. Further controls beyond stratification are used.

CPS. See *Current Population Survey.*

Current Population Survey. A household sample survey conducted monthly by the Bureau of the Census to provide estimates of employment, unemployment, and other characteristics of the general labor force, of the population as a whole, and of various subgroups of the population.

CV. See *coefficient of variation.*

data editing. The process by which assessment responses and other information are verified.

data entry. The process by which assessment responses and other information are transferred from paper to computer.

degrees of freedom. [of a variance estimator]. The number of independent pieces of information used to generate a variance estimate. For the jackknife variance estimator used in Year 15 NAEP, this is at most 32, the number of PSU pairs.

demographic subgroups. See *reporting subgroups.*

derived variables. Subgroup data that were not obtained directly from assessment responses, but through procedures of interpretation, classification or calculation. See also *reporting subgroups.*

design effects. The ratio of the variance for the sample design to the variance for a simple random sample of the same size.

distractor. An incorrect response choice included in a multiple-choice exercise.

District Supervisor. One of 16 supervisors responsible for contacting schools, arranging and conducting introductory meetings, recruiting, training and providing support to Exercise Administrators, distributing and collecting questionnaires, completing administrative reporting forms, and packing and shipping all materials to ETS.

double-length block. A group of assessment exercises, 28 minutes long, created to accommodate the use of longer exercises; used in UBIB spiral administration.

EA. See *Exercise Administrator.*

ECS. See *Education Commission of the States.*

Education Commission of the States. The NAEP grantee prior to Year 15.

Educational Testing Service. The NAEP grantee for Year 15.

entry mode. Processing option under the data entry system: used for the initial transcription of assessment data.

ETS. See *Educational Testing Service.*

examinee. Same as *respondent.*

Excluded Student Questionnaire. An instrument used in the Year 15 assessment; completed for every student who was sampled but was excluded from the assessment.

excluded students. Sample students who were determined by the school to be unable to paticipate because they had limited English-speaking ability, were educable mentally retarded, or functionally disabled.

Exercise Administrator. The person whose primary function was to administer the assessment booklets to the sample students.

exercise. A task designed to measure an objective. Because NAEP does not administer "tests," but instead describes educational achievement over time, the term "exercise" is often used instead of the term "item" or "test item." The terms "item" and "exercise" are used synonymously in this report.

exercise booklet. See *booklet.*

exercise part. Each portion of an exercise that asks a separate question. Parts may all pertain to one stimulus, such as a graph or a table, or may concern the same topic.

exercise pool. The entire set of exercises prepared for a learning area. This set includes recycled exercises developed for previous assessments but not used due to exercise booklet or budgetary constraints and newly developed exercises.

expected value. The average of the sample estimates given by an estimator over all possible samples. If the estimator is unbiased, then its expected value will equal the population value being estimated.

extra subsampling. Subsampling to obtain smaller than desired subsampling fractions. Occasionally used in schools with an unusually large amount of recent growth in numbers of students in order to reduce workload.

field test. A pretest of exercises to obtain information regarding clarity, difficulty levels, timing, feasibility and special administrative problems needed for revision and selection of exercises to be used in the assessment.

final allocation. Usually determined by rounding or adjusting a preliminary sample allocation to integer numbers. See *allocation.*

first stage sampling unit. See *multi-stage sample design.*

foils. The correct and incorrect response choices included in a multiple-choice exercise.

fourth-stage sampling unit. See *multi-stage sample design.*

free-response item. Same as *open-ended response item.*

grade-eligible. An individual who meets the grade definition for one of the Year 15 National Assessment populations: Grade 4, Grade 8, or Grade 11.

grade/age-eligible. A student who meets the age or grade definition for one of the Year 15 National Assessment populations: Grade 4 or Age 9, Grade 8 or Age 13, Grade 11 or Age 17.

group administered package. A package containing exercises which can be administered to groups of students.

group effect. The difference between the mean for a group and the mean for the nation.

holistic scoring. A method of scoring open-ended response exercises that evaluates a response on the basis of overall impression.

imputation. Prediction of a missing value according to some procedure, using a mathematical model in combination with available information. See *plausible values.*

imputed race/ethnicity. The race/ethnicity of an assessed student, as derived from his or her responses to three particular common background items. A Year 15 *reporting subgroup.*

in-school sample design. Sample design for the National Assessment school survey. See *sample design.*

individual completion rate. Proportion of eligibles in the sample who respond by completing one or more assessment packages.

ineligible. Student who is not eligible for National Assessment because he or she does not satisfy grade or age requirements (see *grade/age-eligible*).

informative writing. A writing objective of the Year 15 assessment; writing that is used to share knowledge and convey messages, instructions and ideas.

intelligent data entry system. A set of computer programs and procedures developed in accordance with the NAEP design to validate, verify, transcribe and check for the reasonableness of available data.

IRT. See *item response theory.*

item. See *exercise.*

item block. See *block.*

item booklet. See *booklet.*

item part. See *exercise part.*

item pool. See *exercise pool.*

item response theory. Test analysis procedures that assume a maethematical model for the probability that given examinee will respond correctly to a given exercise.

jackknife. A procedure to estimate standard errors of percentages and other statistics. Particularly suited to complex sample designs.

learning area. One of the areas assessed by National Assessment, e.g., art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies and writing.

literary writing. In the Year 15 assessment, writing from a basis of experience and imaginative ideas to share experiences and understand the world.

LOGIST. A computer program for estimating item parameters by joint estimation procedures.

machine-readable catalog. Year 15 computer processing control information, IRT parameters, foil codes and labels in a computer-readable format.

major strata. Used to stratify the primary sampling frame within each region. Involves stratification by size of community and degree of ruralness (SDOC).

marginal value. A row or column total, the sum of all cell values in the row or column.

meanparts estimator. Estimates a subgroup average score across a set of items by the average of the subgroup scores for each of the items. Can be extended to any linear estimator.

mechanics scoring. A method of scoring open-ended response exercises that evaluates elements of sentence construction, word choice, spelling, punctuation and capitalization.

modal age. The age of the majority of a group of grade-eligible students: Age 9 for fourth graders, Age 13 for eighth graders and Age 17 for eleventh graders.

modal grade. The grade attended by the majority of a group of age-eligible students: the fourth grade for 9-year-olds, the eighth grade for 13-year-olds and the eleventh grade for 17-year-olds.

mode of administration. The method by which students are administered assessment instruments; in Year 15 the modes of administration were spiralled and taped.

multi-stage sample design. Indicates more than one stage of sampling. An example of three-stage sampling: 1) sample of counties (primary sampling units or PSUs); 2) sample of schools within each sample county; 3) sample of students within each sample school.

multiple matrix sampling. Sampling plan in which different samples of respondents take different samples of items.

multiple-county PSU. A primary sampling unit (PSU) composed of two or more counties.

NIE. National Institute of Education.

nine-year-olds. One of the National Assessment target populations. For Year 15, defined as persons born during calendar year 1974.

non-cognitive assessment. The background questions used to collect information from students about activities, attitudes and demographics.

nonresponse. The failure to obtain responses or measurements for all sample elements.

nonsampling error. A general term applying to all sources of error except sampling error. Includes errors from defects in the sampling frame, response or measurement error, and mistakes in processing the data.

objective. A desirable education goal agreed upon by scholars in the field, educators and concerned lay persons, and established through the consensus approach.

objectives re-development. A review of the learning area objectives following the initial assessment of a learning area; carried out by scholars in the field, educators and concerned lay persons. May result in revision, modification or total rewriting of the learning-area objectives to reflect current curricular goals and emphases.

observational unit. The individual units for which characteristics are observed or measurements are obtained.

observed race/ethnicity. Race/ethnicity of an assessed as perceived by the Exercise Administrator.

OERI. Office for Educational Research and Improvement.

OMB. Office of Management and Budget.

open-ended response item. A non-multiple-choice exercise that requires some type of written or oral response.

oversampling. Deliberately sampling a portion of the population at a higher rate than the remainder of the population.

paced tape. A tape recording accompanying each tape administration package to assure uniformity in administration. Instructions are played back from the tape recording to prevent reading difficulties from interfering with an individual's ability to respond. Includes response time.

parental education. The level of education of the mother and father of an assessed student as derived from the student's response to two assessment items. A Year 15 *reporting subgroup.*

participant. See *respondent.*

percent-correct. The estimated proportion of a target population who would answer a particular exercise correctly.

persuasive writing. A writing objective of the Year 15 assessment. Writing that attempts to breing about some action or change.

plausible values. Proficiency values drawn at random from a conditional distribution of a NAEP respondent given his or her response to cognitive exercises and a specified subset of background variables (conditioning variables). The selection of a plausible value is a form of *imputation.*

population. An aggregate of elements, usually individual units with associated characteristics for observation or measurement.

post-stratification. Classification and weighting of selected sampling units by a set of strata definitions after the sample has been selected.

PPS. Probability Proportional to Size.

precision. The expected difference between the expected value and the sample estimate of a population value, as measured by the sampling error.

Primary Sampling Unit. A primary sampling unit. This is the basic geographic sampling unit for National Assessment. A PSU is either a single county or a set of contiguous counties. See also *multi-stage sample design.*

primary trait scoring. A method of scoring open-ended response exercises by evaluating the ability to write for precisely defined purposes. Criteria for evaluating responses are associated with specific point scores in a scoring guide.

Principal Questionnaire. A data collection form given to school principals prior to assessments. The principals respond to questions concerning enrollments, size of the community, occupational composition of the community, etc.

Probability Proportional to Estimated Size (PPES). Selection method where probabilities of selection for sampling units are assigned in proportion to the magnitude of the estimated size measure for each unit.

Probability Sample. A sample in which every element of the population has a known, non-zero probability of being selected.

proportional allocation. Allocation of a sample to strata in proportion to observational units in each stratum.

pseudo-replicate. The value of a statistic based on an altered sample. Used by the jackknife variance estimator.

PSU. See *primary sampling unit.*

public-use data tapes. Computer tapes containing respondent-level cognitive item, background and attitude and demographic data. Available for use by researchers wishing to do secondary analyses of NAEP data.

PUDT. See *public-use data tapes.*

QED. Quality Education Data, Inc. A suplier of lists of schools and school districts with school data for Year 15.

random variable. A variable which takes on any value of a specified set with a particular probability.

reading proficiency scale. Scale (0 to 500) based on IRT upon which levels of reading performance can be measured.

receipt control. Procedures used by scoring staff to check in and screen field materials. Information from these procedures is relayed to assessment administrative staff so that any errors may be corrected.

recycled exercises. The set of exercises that is kept secure from one assessment to the next that will be used to measure changes (growth, stability or decline) in performance for the learning area.

region. One of four geographical regions used in gathering and reporting data: Northeast, Southeast, Central and West (as defined by the Office of Business Economics, U. S. Department of Commerce). A Year 15 *reporting subgroup.*

released item. An item for which results and item text have been reported to the public.

reliability check. The scoring of open-ended response items by a second scorer. In Year 15, twenty percent of these items underwent reliability checks.

reporting subgroups. Groups within the national population for which National Assessment data are reported: sex, race/ethnicity, grade, age, level of parental education, region, and size and type of community.

rescore. If an open-ended exercise was scored under different conditions than presently held or if passage of time may affect scoring, responses from an earlier assessment may be rescored at the same time as responses from a later assessment. Responses from an earlier assessment also may be held and not scored so that they can be scored with responses from a later assessment.

Research Triangle Institute. The NAEP survey subcontractor prior to the Year 15 assessment; drew the sample of PSUs and schools for the Year 15 assessment.

resolution mode. Processing option under the data entry system; used for the correction of erroneous or discrepant data values.

respondent. A person who is eligible for National Assessment, is in the sample, and who responds by completing one or more items in an assessment booklet.

response error. The difference between the observed value and the true value for an observational unit.

response experience. Response rates observed in previous surveys which are used for planning purposes.

response options. Different alternatives to a multiple-choice question that can be selected by the respondent.

response rate. Proportion of specified units for which responses or measurements are obtained.

review conference. A conference held to review the objectives of a learning area to assure their acceptance as measures of the objectives by scholars, educators and lay persons or to review exercises for racial, ethnic, social or regional bias.

RP scale. See *reading proficiency scale*

RTI. See *Research Triangle Institute.*

sample design parameter. A population parameter or a survey parameter, such as an expected response rate, used in designing a sample.

sample design. Specifications for selecting a sample plus specifications for processing the sample data to make estimates. See *sampling plan.*

sample size. The number of units in the sample. (See also *average sample size.*)

sample survey. As opposed to a census, a data collection process whereby only a sample of the population is observed or measured.

sample. A portion of a population, or a subset from a set of units, selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire grade/age population but rather selects a representative sample from the grade/age group to answer assessment items.

sampling error. The error in survey estimates that occurs because only a sample of the population is observed. Measured by *standard error and variance.*

sampling frame. The list of sampling units from which the sample is selected.

sampling plan. Set of specifications and procedures used to select a sample. See *sample design.*

School Characteristics and Policy Questionnaire. A five-page questionnaire completed for each school by the principal or other official; used to gather information concerning school administration, staffing patterns, English curriculum and student services.

school district. Administrative unit of the public school system, usually involving a school system under a single district organization.

school response rate. The response rate for a sample of schools. (See *response rate.*)

scoring guide. A guide for hand scoring an open-ended response item that specifies descriptive or diagnostic categories by giving definitions and example responses.

second-stage sampling unit. See *multi-stage sample design.*

Secondary Sampling Unit (SSU). See *multi-stage sample design.*

secondary traits. Characteristics of a response to an open-ended exercise indicating the presence or absence of elements that are of special significance to the exercise.

secure items. Items not release for public use, in order to be readministered in subsequent assessments to determine whether performance levels have increased, decreased or remained the same.

selection probability. The probability, or chance, that a particular sampling unit has of being selected in the sample.

SES. See *socioeconomic status.*

session. See *assessment session.*

seventeen-year-old. One of the National Assessment target populations. For Year 15, defined as persons born from October 1, 1966 to September 30, 1967.

sex. One of the NAEP *reporting subgroups.* Assessment results are reported for males and females.

simple random sample. Process for selecting n sampling units from a population of N sampling units so that each sampling unit has an equal chance of being in the sample and every combination of n sampling units has the same chance of being in the sample chosen.

single-length block. A group of assessment items, 14 minutes long, containing an average of 12 minutes of reading and writing exercises and two minutes of background and attitude questions.

Size and Type of Community (STOC). One of the NAEP *reporting subgroups*, dividing the communities in the nation into seven groups based on size and other characteristics.

size measure. Value of a variable used to determine the allocation of the sample to strata or used to assign selection probabilities to sampling units within a stratum.

size stratum. A stratum based upon the value of the size measures for units placed in the same stratum; e.g., a stratum for the largest units.

SMSA. See *standard metropolitan statistical area.*

Socioeconomic Status (SES). For sampling, the lower SES portion of the population (approximately 20 percent) is considered a subpopulation to be sampled.

SSU size measure. Measure of size for a secondary sampling unit (SSU).

standard error. A measure of sampling variability for a statistic. Because of NAEP's complex sample design, standard errors are estimated by jackknifing the samples from first-stage sample estimates.

Standard Metropolitan Statistical Area (SMSA). An area defined by the federal government for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an SMSA contains a city of at least 50,000 population plus adjacent areas.

stem. The portion of an item that states the problem or asks the question.

stimulus. For reading items, a visual stimulus used as part of the stem.

STOC. See *size and type of community.*

stratification. The division of a population into parts, called strata.

stratified sample. A sample selected from a population which has been stratified with a sample selected independently in each stratum. The strata are defined for the purpose of reducing sampling error.

student frame. List of grade/age-eligible students within a sample school.

student ID number. A unique identification number assigned to each respondent to preserve his or her anonymity. NAEP does not record the names of any respondents.

student response rate. The response rate for a sample of students. See *response rate.*

study skill item. An item requiring a special learned skill beyond the facility of recognizing and understanding the printed word; for example, the interpretation of a bar graph, telephone bill or table of contents.

subgroup. See *reporting subgroup.*

subject areas. See *learning areas.*

subpopulation. See *reporting subgroup.*

subsampling. Selection of a sample from a larger sample. Also used to describe multi-stage sampling.

subsegmenting. Operation of subdividing the area of a segment into several subareas and selecting one of the subareas.

survey design. All specifications and procedures involved in a survey.

survey population. The population actually surveyed or represented by the sample. May differ from the target population.

systematic sample (systematic random sample). A sample selected by a systematic method; for example, when units are selected from a list at equally spaced intervals.

TAC. See *Technical Advisory Committee.*

tapescript. A script prepared for the announcer to use in producing the paced tape, indicating exactly what is to be read or not read aloud to the students as well as the amount of response time allowed for each exercise. See *paced tape.*

target population. Same as *population.*

Teacher Questionnaire. A nine-page questionnaire completed by selected English and Language Arts teachers; used to gather information concerning year of teaching experience, frequency of writing assignments, teaching materials used, and the availability and use of computers.

Technical Advisory Committee. Committee of experts in areas of educational policy and procedures, mathematics, and measurement theory; provides advice and recommendations concerning NAEP staff technical plans such as sampling, program implementation and analyses.

theta scale. A rescaling of the Reading Proficiency scale that standardizes the combined age and grade samples. Item response theory calculations are carried out in the theta scale for mathematical convenience, then transformed to the Reading Proficiency scale for reporting purposes.

third-stage sampling unit. See *multi-stage sample design.*

thirteen-year-olds. One of the National Assessment target populations. For Year 15, defined as persons born during calendar year 1970.

T-unit. Used to assess the quality of syntax used in an essay; an independent clause and all of its modifying words, phrases and clauses.

UBIB (Unbalanced Incomplete Block) spiralling. Refers to a portion of the spiral design in which each booklet contains a common block of background questions, a single-length block of assessment exercises, and a double-length block of assessment exercises.

unequal probability sampling. A sample selection procedure in which the sampling units have assigned selection probabilities which are not equal for all units.

user tapes. See *public-use data tapes.*

variance. The square of the standard error; the average of the squared deviations of a random variable from the expected value of the variable.

verification mode. Processing option under the data entry system; used for the substantiation of data values.

WARM. See *Weighted Average Response Method.*

weight. A multiplicative factor equal to the reciprocal of the probability of a respondent being selected for assessment with adjustment for nonresponse and perhaps also for post-stratification; an estimate of the number of persons in the population represented by a respondent in the sample. The sum of weights for all respondents at an age level is an estimate of the number of persons in the country at that age level.

Weighted Average Response Method (WARM). A generalization of the Average Response Method allowing the estimation of weighted averages.

Westat, Inc. The NAEP survey subcontractor for the Year 15 assessment.

Winsorizing. Replacement of data values which are more extreme than a given threshold by that threshold value. Bounds the influence of extreme data values on an estimator while maintaining information on the sign of the values.

writing scale. Scale based on *Average Response Method* upon which levels of writing performance can be measured.

Year 01, 02, 03...15. A sequential number assigned to each period of assessment activities in the field. Year 15 pre-assessment activities began in May 1983; assessment activities began in August 1983 and ended in May 1984.

LIST OF REFERENCES

IMPLEMENTING THE NEW DESIGN:
THE NAEP 1983-84 TECHNICAL REPORT

LIST OF REFERENCES

American Psychological Association (1985). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986a). *Writing: Trends across the decade, 1974-1984.* (NAEP Report 15-W-01). Princeton, NJ: Educational Testing Service.

Applebee, A. N., Langer, J A., & Mullis, I. V. S. (1986b). *The writing report card: Writing achievement in American schools, 1984.* Educational Testing Service.

Barone, J. L., Norris, N. A., & Rogers, A. M. (1986). *NAEP 1983-84 public-use data tapes version 3.1 users' guide.* Princeton, NJ: Educational Testing Service.

Beall, G. (1971). *Change-over experiments in practice.* (ETS Research Report RB71-38). Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1964). *The use of special matrix operators in statistical calculus.* (ETS Research Bulletin 64-51). Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1973). F4STAT statistical system. *Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface* (pp. 279-282). Ames, IA: Iowa State University.

Beaton, A. E. (1984). *Statistical issues in data analysis for the National Assessment of Educational Progress.* Paper presented at the annual meeting of the American Statistical Association, Philadelphia, August 1984.

Beaton, A. E. (1985). *NAEP analysis procedures and methodology.* Paper presented at the annual joint meeting of the AERA and NCME, Chicago, April 1985.

Beaton, A. E. (1986). *Behanc, a program to assist in behavioral anchoring* [Computer program]. Princeton, NJ: Educational Testing Service.

Beaton, A.E., Mislevy, R. J., Kaplan, B. & Sheehan, K. M. (1986). *Estimating group effects from sparse, fallible assessment data: Procedures and methodology.* Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283-296.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R. D., and Muraki, E. (1985). *Full-information item factor analysis* (MRC Report No., 85-1). Chicago: National Opinion Research Center.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179-197.

Bock, D. R., Mislevy, R. J. & Woodson, C. E. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4-11, 16.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items and between tests. *Psychometrika, 26*, 347-372.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (eds.), *Principals of modern psychological measurement.* Hillsdale, NJ: Erlbaum.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5-32.

Chromy, J. R. (1979). Sequential sample selection methods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-406.

Cochran, W. F. (1977). *Sampling techniques.* (3rd ed.). New York: John Wiley & Sons.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. (1966). *Equality of education.* SDC No. FS5.238:38001. Washington, DC: National Center for Educational Statistics.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating.* (ETS Research Report 85-30.) Princeton, NJ: Educational Testing Service.

Cook, L. L., & Eignor, D. R. (1984). *Assessing the dimensionality of NAEP reading test items: Confirmatory factor analysis of item parcel data.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, April 1984.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189-199.

Education Commission of the States (1980). *Writing achievement, 1969-79, results from the third national writing assessment: Volume 1 - 17-year-olds, volume 2 - 13-year-olds, volume 3 - 9-year-olds.* Denver, CO: National Assessment of Educational Progress.

Educational Testing Service (1983). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Educational Testing Service (1984). *F4STAT, Version 2.7* [Computer program]. Princeton, NJ: Author.

Felleggi, I. P. (1979). Approximate tests of independence and goodness of fit based on stratified multi-stage samples. *Survey Methodology, 4,* 29-56.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

Goldstein, H. and James, A. (1983). Efficient estimation for a multiple matrix sample design. *British Journal of Mathematical and Statistical Psychology, 36,* 167-174.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al., (Eds.) *The prediction of personal adjustment,* (pp. 319-348). New York: Social Science Research Council.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika, 18,* 277-296.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258-348). Glencoe, IL: The Free Press.

Haberman, J. S. (1977). Log-linear models and frequency tables with small expected cells counts. *Annals of Statistics, 5,* 1148-1169.

Haertel, E. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333-346.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953). *Sample survey methods and theory, Volume I.* New York: Wiley.

Hansen, M. H., Tepping, B. S., Lago, J. A. & Burke, J. (1984). *National Assessment of Educational Progress (NAEP)--the sample and data collection design for Year 15.* Paper presented at the meeting of the ASA, 1984.

Harris, C. W. (1962). Some Rao-Guttman relationships. *Psychometrika, 27,* 247-263.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17,* 65-70.

Hertzog, T., & Rubin, D. B. (1983). Using multiple imputation to handle nonresponse in sample surveys. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys, Volume II: Theory and bibliographies.* New York: Academic Press.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

Holland, P. W., & Rosenbaum, P. R. (In press). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics.*

Holland, P. W., & Zwick, R. (1986). *NAEP scaled scores.* Memorandum to A. E. Beaton, February 13, 1986.

Huber, P. J. (1981). *Robust statistics.* New York: Wiley.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Jacklin, C. N. (1979). Epilogue. In M. A. Wittig and A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning.* New York: Academic Press.

Johnson, E. G. & King, B. F. (1986). *Generalized variance functions for a complex sample survey.* (Technical Report No. 87-72). Princeton, NJ: Educational Testing Service.

Jones, L. V., Burton, N. W., & Davenport, E. C. (1982). *Mathematics achievement levels of black and white youth.* (Report No. 165). Chapel Hill, NC: The L. L. Thurstone Psychometric Laboratory.

Jungeblut, A. (1984). *Assessing the dimensionality of NAEP reading test items: Linear factor analysis models.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984.

Kaiser, H. F. (1963). Image analysis. In C. W. Harris (Ed.), *Problems in measuring change,* pp. 156-166. Madison, WI: University of Wisconsin Press.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35,* 401-415.

Kaiser, H. F. & Cerny, B. A. (1978). Pseudo-images and pseudo-anti-images from the pseudo-inverse of a singular correlation matrix. *British Journal of Statistical Psychology, 31*, 99-101.

Kaiser, H. F., & Cerny, B. A. (1979). Factor analysis of the image correlation matrix. *Educational and Psychological Measurement, 39*, 711-714.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kingston, N. M., & Dorans, N. J. (1981). *The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test.* (GRE Board Professional Report 79-12.) Princeton, NJ: Educational Testing Service.

Kirsch, I. S. & Jungeblut, A. (1986). *Literacy: Profiles of America's Young Adults, Final Report.* (Report No. 16-PL-01). Princeton, NJ: National Assessment of Educational Progress.

Kish, L. (1967). *Survey sampling.* New York: John Wiley & Sons.

Kish, L. & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, *36*, 1-22.

Lago, J. A., Burke, J S., Tepping, B. J., & Hansen, M. H. (1985). *Report on sample selection, weighting, and variance estimation: NAEP-Year 15.* Rockville, MD: Westat, Inc.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*, 259-267.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs* (No. 15).

McDonald, R. P. (1983). Exporatory and confirmatory nonlinear common factor analysis. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement.* Hillsdale, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

Messick, S. J. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. J., Beaton, A. E. & Lord, F. M. (1983). *NAEP reconsidered: A new design for a new era.* (NAEP Report 83-1.) Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984a). *GROUP: Estimation of group effects in univariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984b). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Mislevy, R. J. (1985a). *Inferences about latent populations from complex samples.* (ETS Research Report RR-85-41). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1985b). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R. J. (1985c). *RESOLVE: Estimation of latent distributions by the method of maximum likelihood* [Computer program]. Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J. (1986a). *Exploiting auxiliary information about examinees in the estimation of item parameters.* (ETS Research Report RR-86-18). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1986b). *A Bayesian treatment of latent variables in sample surveys.* (ETS Research Report RR-86-1). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1986c). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Mislevy, R. J. (in progress). *Approximating secondary biases in the analysis of Year 15 NAEP reading plausible values.*

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mosenthal, P. B. (1985). *An analysis of NAEP reading assessment items.* Unpublished manuscript, Syracuse University.

Mosteller, F. & Tukey, J. W. (1969). Data analysis, including statistics. In G. Lindzey and E. Aronson (Eds.), *Handbook of Social Psychology* (2nd ed.). Reading, MA: Addison-Wesley.

Mulaik, S. A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Mussen, P. H., Conger, J. J., & Kagan, J. (1969). *Child development and personality.* New York: Harper and Row.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

NAEP: A proposal submitted in response to Grant Announcement No. PA-82-001: Technical application. (1982). Princeton, NJ: Educational Testing Service.

Reading objectives: 1983-84 assessment. (1984). (NAEP Report 15-RL-10). Princeton, NJ: National Assessment of Educational Progress.

The Reading Report Card: Progress toward excellence in our schools. (1985). (NAEP Report 15-R-01). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Report on field operations and data collection activities. NAEP Year 15. (1984). Rockville, MD: Westat, Inc.

Research Triangle Institute. (1982). *Year 13 field operations and data collection activities, National Assessment of Educational Progress.* Research Triangle Park, NC: Author.

Rivera, C. & Pennock-Roman, M. (1985). *A comparison of race/ethnicity identification methods in the National Assessments of Educational Progress.* Paper presented at the annual meeting of the AERA, 1985.

Rosenbaum, P. R. (1984a). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Rosenbaum, P. R. (1984b). Are the item responses of two groups of examinees consistent with a difference in the distribution of a unidimensional latent variable? (Program Statistics Research Technical Report No. 84-51). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1977). Formalizing subjective notions about the effects of nonresponse in sample surveys. *Journal of the American Statistical Association, 71*, 538-543.

Rubin, D. B. (1978). Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2*, 110-114.

Shah, B. V., Holt, M. M. & Folsom, R. E. (1977). *Inference about regression models from sample survey data.* Research Triangle Park, NC: Research Triangle Institute.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Stout, W.F. (1984). *The statistical assessment of latent trait dimensionality in psychological testing.* (ONR Report). Urbana-Champaign, IL: Department of Mathematics, University of Illinois.

Tepping, B. J. & Hansen, M. H. (1984). *Estimation of population covariances.* Unpublished memorandum.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. *Review of Research in Education, 9*, 377-435.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wingersky, B. (1984). Gramianizing matrices. Unpublished memorandum.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Wingersky, M. S. (1984). *MLE-ABIL: Maximum likelihood estimates of ability* [Computer program]. Princeton, NJ: Educational Testing Service.

Wingersky, M. S. (1986). *Joint estimation procedures, Year 15 NAEP.* Internal technical report. Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A. and Lord, F. M. (1982). *LOGIST V user's guide.* Princeton, NJ: Educational Testing Service.

Wilson, D., Wood, R. L., & Gibbons, R. (1983). *TESTFACT: Test scoring and item factor analysis* [Computer program]. Chicago: Scientific Software.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

Writing objectives: 1983-84 assessment. (1982). (NAEP Report 15-W-10). Princeton, NJ: National Assessment of Educational Progress.

Zwick, R. (1985). *Bias of variance and covariance estimates in NAEP.* Unpublished memorandum.

Zwick, R. (1986a). *Assessment of the dimensionality of the NAEP year 15 reading data.* (ETS Research Report RR-86-4). Princeton, NJ: Educational Testing Service.

Zwick, R. (1986b). *Effects of reader knowledge of student age on NAEP writing scores.* Internal technical report. Princeton, NJ: Educational Testing Service.

SUBJECT INDEX

IMPLEMENTING THE NEW DESIGN: THE NAEP 1983-84 TECHNICAL REPORT

SUBJECT INDEX

Almanacs, 13.4
  format, 13.4.2
  interpretation of, 13.4.3
  types, 13.4.2.1-13.4.2.8

Anchoring scale points, 10.5.2

ARM. *See* Average response method

Assessment design, 2.1-2.9
  assigning items to students, 2.5, 5-5.8
  considerations, 5.1
  database creation, 2.8, 8, 8.6
  data entry system, 8.3
  editing data, 8.4
  field administration, 2.7, 7-7.4
  instruments and items, 2.6, 6-6.4
  processing assessment material, 8, 8.0.1, 8.1
  professional scoring, 8.2
  public-use data tape construction, 8.7
  quality control, 8.5
  reading assessment development, 3, 3.2
  sample selection and instrument collection, 2.4, 4-4.8
  tabular summary, 2.9
  writing assessment development, 3, 3.1

Assessment items, 2.6, 6-6.4 *See also* Reading items, Writing items, Background and attitude items.
  assembling into blocks, 6.1.1
  assigning to students, 2.5, 5-5.8

Assigning items to students, 2.5, 5-5.8
  achieved samples, 5.7
  advantages and disadvantages of spiral design, 5.8
  BIB spiral sample, 5.2
  considerations in design, 5.1
  pairings of item blocks, 5.4
  spiral design, 5.5
  tape sample, 5.6
  UBIB spiral sample, 5.3

Average response method, 11.4
  application to writing data, 11.4.4
  bias of estimator, 11.4.3
  method, 11.4.1
  plausible values, 11.4.2

Background and attitude data analysis, 9.3, 12-12.2

Background and attitude items, 6.1.4
  assembly into instruments, 6.1, 6.1.4
  data, in almanacs, 13.4.2
  and reporting subgroups, 12-12.2
  specific to writing, 3.1.6
  use in validity assessment, 14.1.2.3-4

Batching effect, 11.1.2

Bias, item
  prior knowledge, 3.1.4.2
  review of reading items for, 3.2.3.4

Bias, statistical
  of ARM estimator, 11.4.3
  effects of specification errors on plausible values, 10.3.5

BIB/pace equating, 10.3.6

BIB spiralling. *See also* BIB/UBIB spiralling
  design, 5.2
  impact on dimensionality analyses, 10.1.4

BIB/UBIB spiralling, 2.5, 5-5.8
  advantages and disadvantages, 5.8
  BIB sample, 5.2
  booklets, 2.5.1
  pairings of item blocks, 5.4
  UBIB sample, 5.3

BILOG, 10.3

- Blocks
  - assembly of items into, 6.1
  - pairings in spiral design, 5.4

- Booklets
  - assembly of blocks into, 5.2, 5.3, 6.1
  - spiral, 2.5.1
  - tape, 2.5.2

- Catalog
  - data file, 8.7.5
  - machine-readable, 8.7.8
  - master, 8.6.3

- Clustering, accounting for effects of, 13.2.2

- Cognitive items. *See* Reading items, Writing items

- Construct validity, 14.1.2

- Consultants
  - reading items, 3.2.5
  - writing items, 3.1.8

- Content validity, 14.1.1

- Data analysis, 9-14.2
  - background and attitude, 12-12.2
  - parameter estimation, 13-13.4
  - reading, 10-10.5
  - supplementary studies, 14-14.2
  - writing, 11.0-11.4

- Database creation, 2.8, 8, 8.0.2, 8.6
  - extraction, 8.6.1
  - file merging, 8.6.2
  - master catalog, 8.6.3

- Data entry, 8.3
  - audit trail processing, 8.3.8
  - data spooling, 8.3.11
  - forms, 8.3.7
  - machine considerations, 8.3.2
  - organization and structure, 8.3.3
  - of professional scoring, 8.2.2.7
  - program structure, 8.3.4
  - of questionnaires, 8.3.9
  - of school worksheet, 8.3.5
  - of student data, 8.3.6
  - system requirements, 8.3.1

- Derived variables, 12.1-12.2
  - age, 12.1.7
  - imputed race/ethnicity, 12.1.3
  - parental education, 12.1.6
  - and reporting subgroups, 12.1
  - WARM, 12.2

- Design effects, 14.2
  - from reading assessment, 14.2.1

- Dimensionality, 10.1
  - impact of BIB spiralling on, 10.1.4
  - methods for dichotomous data, 10.1.2
  - methods for reading data, 10.1.3

- Editing data, 8.4
  - professionally scored items, 8.4.2
  - student and questionnaire data, 8.4.1

- Equating, BIB/pace, 10.3.6

- Excluded student
  - questionnaire, 6.2
    - data entry, 8.3.9
    - editing data, 8.4.1
    - quality control, 8.5.2
    - processing, 8.1.5
  - sample, 4.6

- Field administration, 2.7, 7-7.4
  - assessment activities, 7.3
  - assessment materials and forms, 7.4.2
  - conduct of the assessments, 7.3.2
  - drawing the sample, 7.3.1
  - field management, 7.4
  - monitoring field activities, 7.4.1
  - pre-assessment activities, 7.2
  - pre-assessment materials and forms, 7.4.3
  - quality control and evaluation, 7.3.5
  - schedule of field activities, 7.1
  - soliciting cooperation of districts and schools, 7.2.3
  - students sampled, invited and assessed, 7.3.3
  - training district supervisors, 7.2.2
  - training exercise administrators, 7.2.4

- Field testing of reading items, 3.2.3.5, 3.2.3.7

Forms
  for data entry, 8.3.7
  and materials for assessment, 7.4.2
  and materials for pre-assessment, 7.4.3

F4STAT, 9.4

Grade/age, reporting subgroup definition, 12.1.7

Imputation, estimation of variability due to, 13.3. *See also* Plausible values

Instruments, 2.9.1, 2.6, 6, 6.1 *See also* Reading items, Writing items, Background and attitude items, Questionnaires
  development, 2.3
  processing, 8.1.4

IRT. *See* Item response theory

Item response theory, 10.0.1
  and educational assessment, 10.0.2

Jackknife
  alternative estimators, 13.2.5
  estimation of variability, 13.2.3

Joint estimation procedures, 10.2
  item parameter calibration, 10.2.2
  maximum likelihood estimates of proficiency, 10.2.3
  method, 10.2.1

LOGIST, 10.2

Marginal estimation procedures, 10.3
  BIB/pace equating, 10.3.6
  conditional effects estimate, 10.3.3
  effects of specification errors on plausible values, 10.3.5
  general model, 10.3.1
  generation of plausible values, 10.3.4
  item parameter estimation, 10.3.2
  plausible values, 10.3.1.1

Mode of administration, effect on estimates of writing performance, 11.2

Non-cognitive items. *See* Background and attitude items

Nonresponse, adjustment for, 13.1.2
  school, 13.1.3
  student, 13.1.4
  students in spiral sessions, 13.1.6
  students in tape sessions, 13.1.5

Objectives
  reading, 3.2.1
    development, 3.2.2
  writing, 3.1.1
    development, 3.1.2

Parameter estimation, 9.4, 13-13.4
  uncertainty due to sampling variability, 13.2
  variability due to imputation, 13.3
  use of almanacs, 13.4
  weighting procedures, 13.1

Parental education, reporting subgroup definition, 12.1.6

Participation results, student, 4.7

Plausible values, 10.3.1.1
  and average response method, 11.4.2
  effects of specification errors on, 10.3.5
  generation of
    in reading analysis, 10.3.4
    in reading trend analysis, 10.4.5

Post-stratification, 13.1.9

Pre-assessment
  activities, 7.2
  materials and forms, 7.4.3

Primary sampling units, 4.1
  sample characteristics, 2.9.2

Processing assessment materials, 8, 8.0.1, 8.1
  administration schedules, 8.1.2
  professionally scored items, 8.1.4.2
  questionnaires, 8.1.5
  receipt of materials, 8.1.1
  school worksheets, 8.1.3
  student instruments, 8.1.4

Professional scoring, 8.2
  data entry, 8.2.2.7

Professional scoring *(continued)*
  description, 8.2.1
  editing data, 8.4.2
  operation, 8.2.2
  processing data, 8.1.4.2
  reliability and resolution, 8.2.2.6
  scorers, 8.2.2.1
  training, 8.2.2.2-8.2.2.4
  types
    holistic, 8.2.1.2
    mechanics, 8.2.1.3
    primary trait, 8.2.1.1

Proficiency
  of American students, 15-15.1
  and construct validity, 14.1.2.1
  data, in almanacs, 13.4.2.3-13.4.2.6
  maximum likelihood estimates of, 10.2.3

PSAT scores and construct validity, 14.1.2.3, 14.1.2.4

PSU. *See* Primary sampling units

Public-use data tape construction, 8.7
  codebooks, 8.7.6
  data definition, 8.7.3
  data file catalogs, 8.7.5
  data file layouts, 8.7.4
  file definition, 8.7.1
  machine-readable catalog files, 8.7.8
  SAS and SPSS-X control files, 8.7.7
  variable definition, 8.7.2

Quality control, 8.5
  during field administration, 7.3.5
  of questionnaire data
    excluded student, 8.5.2
    school characteristics & policy, 8.5.4
    teacher, 8.5.3
  of student data, 8.5.1
  summary, 8.5.5

Questionnaires
  data entry, 8.3.9
  editing data, 8.4.1
  excluded student, 6.2
  processing data, 8.1.5
  quality control, 8.5.2-8.5.4
  school characteristics and policy, 6.4
  teacher, 6.3

Race/ethnicity, reporting subgroup definition
  imputed, 12.1.3
  observed, 12.1.2

Reading data analysis, 9.1, 10-10.5
  dimensionality, 10.1
  item response theory, 10.0.1, 10.0.2
  estimation procedures
    joint, 10.2
    marginal, 10.3
  scaling, 10.5
  trend, 10.4

Reading items, 3.2.4, 6.1.2
  assembling into blocks, 6.1.1
  bias review, 3.2.3.4
  data, in almanacs, 13.4.2
  design effects, 14.2.1
  development, 3, 3.2
    consultants, 3.2.5
  field testing, 3.2.3.5, 3.2.3.7
  objectives, 3.2.1
    development, 3.2.2
  and student instruments, 6.1
  use in trend analysis, 10.4.2

Reading scale, 10.5
  anchoring scale points, 10.5.2

Reading trend
  analysis, 10.4
  data, in almanacs, 13.4.2.6-13.4.2.7
  estimation of, 10.4.1
    conditional effects for, 10.4.4
    item parameters for, 10.4.3
  generation of plausible values for, 10.4.5
  selection of items for, 10.4.2

Region, reporting subgroup definition, 12.1.5

Reliability
  inter-rater (writing data), 11.1.1
  professional scoring, 8.2.2.6

Reporting subgroups, 12-12.2
  definition of
    grade/age, 12.1.7
    imputed race/ethnicity, 12.1.3
    observed race/ethnicity, 12.1.2

Reporting subgroups *(continued)*
    parental education, 12.1.6
    region, 12.1.5
    sex, 12.1.1
    size and type of community, 12.1.4
  and derived variables, 12.1

Sample selection, 2.4, 4-4.8
  excluded student sample, 4.6
  primary sampling units
    characteristics, 2.9.2
    sample, 4.1
  schools
    assignment of sessions to, 4.4
    characteristics, 2.9.2
    initial sample, 4.2
    updating the sample, 4.3
  students
    characteristics, 2.9.3
    participation results, 4.7
    samples, 4.5
  tabular summary, 2.9
  teacher
    characteristics, 2.9.2
    sample, 4.8

SAS control files, 8.7.7

Scale. *See* Reading scale, Writing scale

Schedule
  of assessments, 2.2
  of field activities, 7.1

Schools
  assignment of sessions to, 4.4
  characteristics and policy questionnaire, 6.4
    data entry, 8.3.9
    editing data, 8.4.1
    processing data, 8.1.5
    quality control, 8.5.4
  initial sample, 4.2
  sample characteristics, 2.9.2
  updating the sample, 4.3

Sex, reporting subgroup definition, 12.1.1

Size and type of community, reporting subgroup definition, 12.1.4

Spiralling. *See* BIB/UBIB spiralling

SPSS-X control files, 8.7.7

Stratification, accounting for effects of, 13.2.2

Students
  assignment of items to, 5-5.8
  assessment data
    editing, 8.4.1
    entry, 8.3.6
    processing, 8.1.4
    quality control, 8.5.1
  assessment instruments, 6.1
  associated teacher sample, 4.8
  drawing the sample of, 7.3.1
  participation results, 4.7, 7.3.3
  sample characteristics, 2.9.3
  samples of, 4.5

Supplementary studies, 9.5, 14-14.2
  design effects, 14.2
  validity issues, 14.1

Systematic selection, accounting for effects of, 13.2.2

Tape
  booklets, 2.5.2, 5.2, 5.3, 6.1
  sample, 5.6

Teachers
  questionnaire, 6.3
    data entry, 8.3.9
    editing data, 8.4.1
    processing, 8.1.5
    quality control, 8.5.3
  sample, 4.8
  sample characteristics, 2.9.2

Training
  district supervisors, 7.2.2
  exercise administrators, 7.2.4
  professional scorers
    holistic, 8.2.2.3
    mechanics, 8.2.2.4
    primary trait, 8.2.2.2

Trend. *See* Reading trend, Writing trend

UBIB spiral design, 5.3. *See also* BIB/UBIB spiralling

Validity, 14.1
  construct, 14.1.1
  content, 14.1.2
  issues, 14.1
  summary of evidence, 14.1.3

WARM variables, 12.2

Weighted average response method variables, 12.2

Weighting, 13.1
  adjustments
    for missing tape session, 13.1.7
    for nonresponse, 13.1.2
    for school nonresponse, 13.1.3
    for student nonresponse, 13.1.4
    for student nonresponse, spiral sessions, 13.1.6
    for student nonresponse, tape sessions, 13.1.5
  computation of base weight, 13.1.1
  final student weight, 13.1.10
  full-sample weight, 13.1.10
  post-stratification, 13.1.9
  procedures, 13.1
  trimming to reduce mean squared error, 13.1.8

Writing data analysis, 9.2, 11-11.4
  average response method of scaling, 11.4
  batching effect, 11.1.2
  effect of mode of administration, 11.2
  inter-rater reliability, 11.1.1
  trend, 11.3
  writing data, 11.1
    descriptive statistics, 11.1.3

Writing items, 3.1.5, 6.1.3
  assembling into blocks, 6.1.1
  background and attitude, 3.1.6
  data, in almanacs, 13.4.2
  development, 3, 3.1
    consultants, 3.1.8
  objectives, 3.1.1
    development, 3.1.2
  and prior knowledge bias, 3.1.4.2
  and student instruments, 6.1
  and types of writing tasks, 3.1.4.1
  use in trend analysis, 11.3

Writing scale, 11.4. *See also* Average response method

Writing trend
  data, in almanacs, 13.4.2.8
  estimation of, 11.3

9066402 • S37P.6 • 249379 • Printed in U.S.A.